REPORT RESUMES

ED 003 330 24

THE INTEGRATION, STORAGE, AND RETRIEVAL OF BIBLIOGRAPHIC DATA IN ENGLISH STUDIES.

BY- SAWIN, LEWIS AND OTHERS
COLORADO UNIV., BOULDER
REPORT NUMBER CRP-2189 PUB DATE 65
EDRS PRICE MF-$0.18 HC-$4.88 122P.

DESCRIPTORS- *ENGLISH, LITERATURE, *BIBLIOGRAPHY, *SYSTEMS APPROACH, SYSTEMS DEVELOPMENT, COMPUTER PROGRAMS, INFORMATION PROCESSING, INFORMATION SYSTEMS, *COMPUTERS, *RESEARCH TOOLS, INFORMATION RETRIEVAL, BOULDER, COLORADO

THE INTEGRATED BIBLIOGRAPHY PILOT STUDY EXPLORED PROBLEMS INHERENT IN CREATION OF AN INFORMATION SERVICE USING A SINGLE BIBLIOGRAPHICAL COMPILATION CONTAINING EVERY ITEM WHICH HAS EVER BEEN LISTED IN A MAJOR BIBLIOGRAPHY OF ENGLISH STUDIES. PROVISION WAS MADE ALSO FOR CONTINUOUS ADDITION OF NEW ITEMS. A COMPUTER-BASED INFORMATION PROCESSING SYSTEM WAS USED TO (1) INTEGRATE OR MERGE SOURCE BIBLIOGRAPHIES, USING A MINIMUM OF HUMAN EFFORT, (2) STORE THIS INFORMATION, AND (3) RETRIEVE ONE-SUBJECT BIBLIOGRAPHIES. TEST OF THE SYSTEM WITH A DATA BASE FROM THE BIBLIOGRAPHIC CORPUS DEMONSTRATED THE FEASIBILITY OF AN INFORMATION SERVICE FOR ENGLISH STUDIES BASED ON AN INTEGRATED BIBLIOGRAPHY. A MACHINE-READABLE RECORD REPLICATED EACH SOURCE BIBLIOGRAPHY WITHOUT CHANGE OF FORMAT OR OMISSION OF DETAIL. DETAILED FLOW DIAGRAMS AS WELL AS COMPUTER PROGRAMS WERE PREPARED FOR ALL OPERATIONS. RECOMMENDATIONS WERE THAT THE SYSTEM IMPLEMENTATION BE COMPLETED SO THAT (1) PRODUCTION-TYPE OPERATIONS COULD BE CARRIED THROUGH ON RELATIVELY LARGE QUANTITIES OF DATA, (2) RETRIEVAL CAPABILITY COULD BE THOROUGHLY TESTED, AND (3) COST ESTIMATES COULD BE OBTAINED FOR OPERATION OF THE SYSTEM AS AN INFORMATION SERVICE. (JM)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

THE INTEGRATION, STORAGE, AND RETRIEVAL OF BIBLIOGRAPHIC DATA IN ENGLISH STUDIES

Cooperative Research Project #2189

Lewis Sawin
Charles Nilon
(Co-Directors)
and
Roger Clark

University of Colorado
Boulder, Colorado

1965

The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare.

ED 003330

ACKNOWLEDGMENTS

The co-directors would like to express their deep gratitude to all those, both on and off the IBPS staff, who helped make this project a success: to Jack McCormick, Librarian, National Center for Atmospheric Research, whose generous donation of energies and facilities quite literally made the implementation of the system possible; to Lawrence Thorsheim and his colleagues at the U.S. Air Force Finance and Accounting Center, whose counsel and advice kept us steering a reasonably straight course; to Jessica Melton, Center for Documentation and Communication Research, who put us on that course in the first place and has encouraged our progress; to Harry Arader and Steve Furth of IBM New York, Frank B. Rogers of the University of Colorado Medical School, and Walter MacIntyre, Director of the University of Colorado Graduate School Computing Center, for their well-considered advice and opinions; and to the members of the Committee on an Integrated Bibliography, Lewis Leary chairman, for their unflagging interest and encouragement.

Finally, our thanks to Curtis D. Benster, Technical Director, and William S. Crosby, IBPS System Analyst, who devised and implemented a system far exceeding in sophistication and application our most extravagant expectations; and to Edward DeAvila for his faultless draftsmanship. Special thanks are also due to our graduate research assistants, Robert Cherry, Eugene Cunnar, Allen R. Penner, Gordon Eriksen, and R. Phillip Loney; and to John Hummel of Dura Business Machines of Colorado.

Assembled program listings for all computer programs described in this report, and computer-independent analytic diagrams for all but certain diagnostic output programs, are on file on microfilm with the U. S. Office of Education.

TABLE OF CONTENTS

I. INTRODUCTION 1

A. The Integrated Bibliography Concept 1

B. The Integrated Bibliography Pilot Study (IBPS) . 5

II. SYSTEM DESCRIPTION 7

A. Summary 7

B. Implications of the Four Principles of the System Design 8

C. Input Functions 19

D. Technical Aspects of the System 25

1. The Work to be Accomplished 25
2. Segmentation of the System 30
3. Evolution of Programs 33
4. Description of Programs 37

E. Retrieval 53

III. CONCLUSION 66

APPENDICES

A. Diagrams 69

IBPS System Design
Legend to Programs
MASC Phase I
MASC Phase II
MASC with SLASH Subroutine

B. "The Integrated Bibliography Pilot Study--Technical Rationale," Curtis D. Benster, Technical Director 70

C. The 1964 IBPS Conference 76

D. Special Studies 83

- Indexing Questionnaire 84
- Indexing Studies 86
- "A Selected Bibliography on Information Retrieval" 87
- "A Bibliography of Bibliographies of American Literature". 89
- Cost Estimates 92

E. Glossary 93

I. *INTRODUCTION*

A. THE INTEGRATED BIBLIOGRAPHY CONCEPT.

English Studies, like other disciplines, faces an over-information crisis. Current publication in the field--perhaps 12,000 items in a year--is dispersed in thousands of journals and books, and every year this volume increases. Furthermore, this crisis has a historical dimension: in English Studies current publication does not necessarily render obsolete the scholarship and criticism of the past; awareness of current scholarship is, quite literally, only the beginning of knowledge. Control over this bibliographical corpus is woefully inadequate by the standards of modern information science, despite the existence of a very large number of enumerative bibliographies, both current and retrospective, serial and self-contained (or terminal), which report the existence of published (and in some instances unpublished) scholarship and criticism. These range from one-subject bibliographies (e.g., of studies relating to one author) through bibliographies of studies relating to subjects within a single literary period or genre, to all-field compilations.

Two features of this profusion of bibliographical sources are important both to an understanding of the nature of the bibliographical problem and to its solution: (1) the substantial duplication among compilations, and (2) the differences among the bibliographies. Although a number of bibliographies cover the same subjects and report the work of the same years, their coverage is best described as overlapping, rather than congruent, which they certainly are not. No single bibliography is comprehensive (all fields, all relevant subjects, all years in which scholarship and criticism were produced, all published and unpublished material in every subject area). Each is limited; each differs from the others in some or all of five different ways: (1) the span of years covered; (2) the principle of selectivity governing the inclusion or exclusion of items from the compilation; (3) the completeness and accuracy of individual entries in the compilation; (4) the degree of annotation (if any) provided with individual entries; (5) the stylistic conventions for the expression of a bibliographical citation. Taken together, the range of reportage is impressive; but each bibliography is a separate entity.

Because of these differences, the researcher in English Studies cannot rely on any one, or even on three or four, bibliographical sources when he prepares a thoroughgoing bibliography of studies on any specific subject. Preparation of such a thoroughgoing bibliography on a single subject is achieved only after days or weeks spent in conflating all relevant sources, each of which must be consulted because it may correct errors made in other sources, and contain new entries, new annotations, new details.

The inevitability of this tedious, uncreative, and time-consuming effort necessarily produces some serious side effects, e.g., (1) lack of time to prepare adequate bibliographies cuts off the scholar and teacher from the results of research and the insights of criticism in the special areas of his own research and his day-to-day teaching; (2) such an impediment to communication between the teacher/scholar and research and criticism generates a rationalization that these are less valuable to him than, in fact, they are; (3) owing to widespread ignorance of work done by others, recent scholarship and criticism are needlessly duplicating work already done.

The resources of the total record now spread among many bibliographies can, however, be harnessed to produce an effective information system, through their *integration* into one compilation which, in addition to the advantages of unification, possesses the retrieval capabilities offered by a combination of the electronic computer and modern indexing concepts.

The Integrated Bibliography for English Studies may therefore be defined as a single bibliographic compilation containing every item which has ever been listed in a major bibliography whose subject contents are related to the discipline of English Studies, with provision for continuous addition of new items, the whole indexed and stored in such a way that retrieval of many discrete segments of the total record can be accomplished quickly and accurately on demand. Printed output must conform to accepted typographical and stylistic conventions and be capable of replicating every typographical sophistication found in the sources (e.g., a variety of fonts, foreign-language diacritics and characters, special symbols, and non-Latin alphabets).

Creation of such an Integrated Bibliography will, it is believed, meet the information crisis in English Studies and stimu-

late new interest in scholarly and critical writing--by teachers as well as research scholars.

B. THE INTEGRATED BIBLIOGRAPHY PILOT STUDY (IBPS).

The Integrated Bibliography Pilot Study was designed to be an exploration of the integrated bibliography concept described above. The specific objectives of the Pilot Study may be summarized as follows: (1) to encounter and solve problems inherent in the creation of an Integrated Bibliography for English Studies and (2) to design an information system according to the criteria stated above; (3) to provide a data base on which to test the system by conducting a "trial bore" in the bibliographic corpus and preparing entries from this trial bore for input to the computer.

It was clear from the outset that a systems approach not only was desirable but had been made mandatory by the terms in which the Pilot Study had been defined. Indeed, it would be necessary not only to design a system but also to modify it in terms of problems encountered along the way; to convert theory into practice by writing and testing specific programs for the various computer operations; and to carry out production-type operations on input data, after facing and solving complicated problems relating to the transcription of this data into machine-readable form.

Consultation with Frank B. Rogers, University of Colorado Medical School; Walter MacIntyre, Director, University of Colorado Graduate School Computing Center; Jessica Melton, Center for Documentation and Communications Research, Western Reserve University; and Curtis D. Benster and William S. Crosby of the firm of Linguamation led to the adoption, in June 1964, of a general system design (by William S. Crosby) which emphasized these four principles:

(1) That, in general, the Pilot Study should not undertake to

look behind the source bibliographies to the original documents.

(2) That no scrap of information given by these printed source bibliographies should be discarded or lost.

(3) That the amount of editorial judgment required at the keyboard--and in editing and annotating material for the keyboard--should be held to a minimum.

(4) That the Master File should utilize the American Standard Code for Information Interchange (ASCII).

During the IBPS Conference in July 1964, only one of these principles (#2) escaped sharp criticism. Reflection upon the nature of the criticism (it seemed to be based upon [a] a misunderstanding of the *experimental* nature of the Pilot Study, and [b] distrust of new ideas on the part of the non-literary participants) and further consultation with Harry Arader of IBM, Lawrence Thorsheim of the U.S. Air Force Finance and Accounting Center, Jack McCormick of the National Center for Atmospheric Research, Curtis D. Benster, and William S. Crosby, led to a decision to adhere to the system design, the salient features of which were derived from the four principles listed above.

II. SYSTEM DESCRIPTION

A. SUMMARY.

The system design specifies that the source bibliographies be transcribed *literatim*, either on paper tape or punched cards, with no pre-editing save of transcription accuracy, for input to the computer. The machine-readable record replicates without change or omission the format of each source bibliography--subject-headings and annotations as well as the bibliographic items (Bibliograms) themselves. Furthermore, the typography of all these elements is preserved, through the use of escape operators which expand the capabilities of the input device. The omission of pre-editing functions intentionally throws upon the computer the burden of identifying (through analysis of the signification of typographical gestalts, a process analogous to the intellectual decisions made by a human reader) the subject-headings and annotations as distinct from the Bibliograms proper, and within the Bibliogram, each of its discrete elements (e.g., author, title, publication facts). This replacement of human editing not only looks forward to the day when optical reading devices will be employed to input data to the computer, but also automates the entire process of integration. The resultant Master File preserves all information presented at input. The preservation of all subject-headings and annotations *in relation to each Bibliogram* (e.g., the location of each entry in the hierarchy of subject-headings under which it is placed in each source where it appears) provides an inherent retrieval capability, which may be exploited by coordinate searches of various kinds conducted within and/or without the Bibliogram proper. Print-outs in response to search queries may appear in simple or sophisticated typography, according to the printing device employed or the requirements of the user, in any specifiable bibliographical style, with or without associated predicative matter (subject-headings, annotations, etc.) as desired. Provision is made both for the addition of new bibliographies and for the application of a different indexing system, if desired.

B. IMPLICATIONS OF THE FOUR PRINCIPLES OF THE SYSTEM DESIGN.

1. Adequacy of Source Bibliographies. The decision that the Pilot Study should not undertake to look behind the source bibliographies to the original documents was based on the assumption that data obtainable from the source bibliographies themselves would be sufficient for a more-than-adequate retrieval system if these data were properly manipulated. A second consideration was the cost of the alternative: a system based on an examination and indexing of the contents of the full text of all items in the data base. For a full Integrated Bibliography the cost of such a system would appear to be prohibitive. It seemed more valuable, therefore, to exploit to the utmost the information given by the bibliographies themselves, i.e., obtainable from title, publication facts, annotations and abstracts, and from the placement of each item in the hierarchical category structure employed by each of the various bibliographical sources. Analysis of the *inherent retrieval capability* provided by this information demonstrates that this capability is substantial. However, the possibility that inherent retrieval capability may be supplemented or even superseded by indexing derived from examination of full text is provided for in principle.

2. Preservation of Source Information: The IBPS Escape Symbols. The need to preserve *all* information from the source bibliographies is obvious in light of the decision to exploit inherent retrieval capability. By "*all* information" is meant not only every word, but every detail of typography (fonts, stylistic

details; in sum, the actual physical appearance of the pages of each bibliographical source). Such preservation of all information actually amounts to a replication of the entire format of each source bibliography.

The rule to typists and keypunchers, therefore, has been "follow copy to the letter." All information present in the printed sources is preserved; none is discarded either at the input keyboard, nor at any other point along its route to its final resting-place in the magnetic-taped Master File.

Such an accomplishment, which in theory makes unnecessary forever afterward any reexamination of the source bibliographies for further detail, is a most worthwhile one, but it was not achieved without some difficulty.

The problem lies in the restricted keyboard of the generally available input devices (i.e., tape-punching typewriter or card punch). The meager font of 84 or 88 characters, to which typewriters are limited, imposes severe constraints. The usual method of escaping these constraints is by a set of escape symbols. The IBPS Escape Symbols, however, differ substantially from the sets now in use (e.g., that described by Martin Kay and Theodore Ziehe of the RAND Corporation[1]) not only in that they are mnemonic and hence easily manageable by the at-present all-too-human keyboard operator, but in that they require no discrimination at the keyboard between different uses of the same character (e.g., the discrimination between *period* and *decimal point*).

---

[1] *Natural Language in Computer Form* (Santa Monica, Calif.: The RAND Corporation, 1965).

An Escape Symbol, as the term is used here, is a device employed to "escape" the limitations of the typewriter font and to record a requisition for any character from a virtually unlimited font. Escape Symbols are keyed as needed in the course of operating the input typewriter, using none but the ordinary graphic keys, and the record of their employment is verifiable by inspection of the hard copy produced as a by-product of the tape-punching operation. (These symbols appear in computer-processed printouts only when needed for diagnostic purposes.)

Escape Symbols take the form /X/, /XX/, or /XXX/: one, two, or three characters enclosed between a pair of slashes.

It would have been possible to design a far more compact code than this, especially if the project were concerned only with one language, or even with only the languages of Western Europe; but it is to be remembered that the transcribing of bibliographies, where one may encounter items in virtually any language--Czech, Swedish, Spanish, Greek, and Russian--perhaps all on the same page, is rather more challenging than the transcribing of a long text in a single language. It was felt, therefore, that mnemonic ease (both for the keyboard operation and for any subsequent proofreading) was mandatory, a requirement paramount to any economy in the mere number of keystrokes.

The Symbols are of three different genera, which have been given the names "Shifter," "Modifier," and "Inserter."

A *Shifter* (for example /bf/ = bold face) orders, for every ensuing graphic, a shift from the standard character-set into a specified different set, the shift remaining in effect until a

Countershifter character is written.

The Countershifter character chosen was the dollar sign. Not only does its "crossed-out-S" appearance suggest its Countershifter function, but also it was expected that the occurrence of a dollar-sign-per-se in text would seldom need special precautions against being misconstrued, for a textual dollar sign would rarely fail to be preceded by a space or equivalent, while a Countershifter ought never to be preceded by anything other than a non-space graphic--the last of the characters affected by the Shifter.

In the event of a cumulation of Shifters (e.g., /cy//bf/ for Cyrillic bold-face) the most recent of the Shifters is understood to be the one cancelled by the Countershifter--"last in, first out."

A *Modifier* (for example /ac/ = acute accent) is an Escape Symbol which, unlike a Shifter, requires no Countershifter to terminate its effect. Depending on which Modifier it is, it is self-terminating at the end of its allotted span of characters--most commonly, at the end of the next single character only. The most frequently occurring type of Modifier is the one which affixes a diacritic mark to an otherwise conventional letter.

With respect particularly to the "diacritic" Modifiers, the question arose initially as to which of two possible sequences was best: diacritic first, and then main letter, or the other way around. It was decided to adopt the former sequence, to which a typist used to a typewriter having diacritics on "dead" keys would more readily adjust.

An *Inserter* (for example, /da/ "dagger"), unlike a Shifter or a Modifier, is completely self-contained. It stands for the insertion in its own place of a specified character.

The classic input medium for computers is the *punched card*, and it was considered that to be worthy of the name the Pilot Study should employ no system which should not be applicable to a wide range of input media and equipment other than those of our own preference. Consequently a portion of the data has been keypunched.

When the card punch is used as input device rather than the tape-punching typewriter some additional measures are necessary for transcribing typographical detail. The keyboard of a typical card punch is more severely restricted even than that of a typewriter, its font comprising a meager 47 characters. For input transcribed from printed copy onto punched cards the following additional conventions have been employed:

(1) Every capital letter is preceded by an asterisk.

(2) A prefixed asterisk is used also to transform comma, period, and apostrophe respectively into semicolon, colon, and exclamation point, and for certain other punctuation transformations.

(3) But a question mark is represented by a plus sign, and quotation marks by an equals sign; dashes of various lengths are improvised as reiterated hyphens, as in typescript.

(4) An additional Shifter, /it/ (= italics) is employed, since the card punch has no underline.

Both typist and keypuncher have been provided means of ensuring that whenever a character like *, $, or / is to be construed literally rather than as part of the system of Escape Symbols its literal interpretation can be "forced." The IBPS code-transformation programs, however, are capable of making the discrimination in most cases.

The Escape Symbols are interpreted by a dictionary of IBPS Escape Symbols (DICTES) which is a part of the program MASC (Mach code to ASCII) and which is in a form sufficiently extensive to meet the requirements presented by the bibliographies now being processed. (DICTES will, for example, cope with most diacritic modifications of the ordinary Latin alphabet, as well as with a variety of type fonts.) DICTES is a selection from a much larger list of IBPS Escape Symbols. That list is susceptible to almost indefinite augmentation, and was developed in order that the more demanding requirements which may be placed upon it in the future should not force any drastic overhaul of DICTES.

The use of the IBPS Escape Symbols makes possible not only a full exploitation of inherent retrieval capabilities, but also the production, when and if appropriate, of a printed book of elegant typographical appearance via type-composing machines such as the paper-tape controlled Linotype, Linofilm, or Photon machines. (Two other forms of output are available from the system and will undoubtedly be used for other kinds of production output: that obtained from a high-speed line-printer--with Escape Symbols suppressed--and the more slowly produced but more attractive output available from a paper-tape controlled typewriter.)

3. Minimization of Editing. The decision to minimize human editorial functions deliberately throws upon the computer the burden of carrying out these functions. This decision, which made

the machine-handling operations vastly more complex than had been anticipated when the proposal for the Pilot Study was written, has proved to be a sound one. It is an important innovation, whose success in this project marks, it is believed, a significant breakthrough in language data processing. The computer accepts essentially unedited material diverse in style and content. Through programs which will be described below, the computer is made to carry out editorial functions needed to provide analysis of the Bibliograms and predicative matter adequate to the *integration* of these data into a Master File and to the development of the inherent retrieval capability of these data. (For example, the computer, working with data transcribed *literatim* from the sources, determines such points in the Bibliogram as where "Author"[1] ends and "Title of Article" begins, what the boundaries of "Title of Journal" are, whether the Bibliogram describes a book, an article, or a *Festschrift* essay, and many more.)

Rather than debate the alternative--employing the capability of a properly-trained keypuncher or typist and/or a staff of editors to restructure the raw data to conform to one style and delimit the boundaries of various kinds of information within and without each entry before or during the transcription process to a machine-readable medium--what is important to realize is that all visual-manual methods of transcription are to become obsolescent a few years hence with the improvement of devices for automatic optical scanning, and that this Pilot Study should prepare for that revolution. No pre-editing should be demanded of our human transcribers/editors against the day when we shall be unable to demand any at all from the machines that will have supplanted them.

---

[1] Because a bibliography of secondary sources in English Studies includes a high percentage of items which are about *writers* of one sort or another, a certain ambiguity attaches to the use of such words as *author*, *writer*, etc. The following terminology has therefore been adopted:

(a) *Author* shall describe the creator of an item (book, article, etc.) listed in a bibliography.

(b) *Surname subject* shall describe the name of a person, writer or otherwise, who is the subject of any such item.

(c) *Surname subject-heading* shall describe a subject-heading (given in a source bibliography) which is the name of a surname subject.

For example, A.C. Howell is the *author* of an item ("Walt Whitman, Singer of the American Spirit") which includes the *surname subject* Walt Whitman, and which is indexed (in *PMLA*, "Bibliography for 1958--American") under the *surname subject-heading* WHITMAN.

(a) Integration. To produce a merged or "integrated" bibliography it is necessary to identify each entry in each source bibliography as (a) *unique*--that is, appearing only once in the entire data base--or (b) *duplicative*--that is, appearing in two or more bibliographies in the data base. However, *duplicative* entries are seldom if ever *identical*, for two reasons:

1. The occurrence of errors, e.g., incorrect spelling of words, incorrect report of pagination, in one or more of the duplicates. The error may be mechanically or humanly generated; the error may have been made in the source bibliography, or by project staff in transcription from source to paper tape. In any case, error is a factor in reducing the possibility of *identity* of *duplicates*.

2. Variations among bibliographies both in style and content. A bibliographical entry gives several different kinds of information; that is, it contains several "elements," e.g.,

author, title, publication data, reviews and annotations. Each source bibliography consistently employs one style for expressing information, but each style differs from every other in some or all of the following ways:

a. number of elements

b. sequence of elements

c. typographical details.

The IBPS program CNFORM analyzes the structure of a bibliographical entry (*any* bibliographical entry), identifying the elements so that they may be compared with the same elements of other entries (regardless of their order within the entry), and duplicative Bibliograms identified as such and merged with each other.

(b) *Retrieval*. Searches will examine both the words and content-elements of Bibliograms, *and* the predicative matter, which includes annotations and abstracts on the one hand, and the hierarchical structure of categories in which the merged Bibliogram was originally found in *all* of the sources where it appeared on the other. Search of the Bibliogram proper will be speeded by the insertion of element-identifying "information separators" within the Bibliogram; searches involving the predicative matter cannot take place unless both Bibliograms and predicative matter are correctly identified and delimited by information separators. It will be remembered that the "follow copy" rule for typists preserves intact, and in the right order, the original Bibliogram, *and* its location in every hierarchy of subject categories.

4. The American Standard Code for Information Interchange (ASCII). The American Standard Code for Information Interchange

is admirably suited to assist in the solution of the problems of transcription, integration, and retrieval.

Many different codes are used by various input/output devices and even within computers themselves. But it is anticipated that within a few years' time they will be superseded by ASCII. This code was therefore adopted as a common language for storage. This is a pioneer utilization of a code which, in addition to being a standard, offers several other advantages:

(1) It furnishes a *definitive list* of the characters that can be stored directly, each in a one-character bit pattern (as distinct from all other characters outside the list, to store which requires the use of escape symbols or other artifices). For example, it is reassuring to know that as ASCII comes progressively into wider use such questions as whether a given piece of equipment will or will not recognize or produce square brackets will be less and less a matter of conjecture.

(2) In the design of ASCII great care has been bestowed on anticipating problems of data-processing "collation" (i.e., arranging items in alphabetical order) and simplifying their solution.

(3) The American Standard Code for Information Interchange, X3.4-1963, approved as American Standard by the American Standards Association on June 17, 1963, includes eight hierarchically ordered "information separators" in addition to the ordinary inter-word "space." X3.4-1963 is at present undergoing revision, however, and it is expected, with a high degree of certainty, that its revised form will include only *four* such separators. Therefore, the

IBPS System has employed only four of the ASCII information separators, though the original eight would have been of immense utility in the segmentation of bibliographic entries into their elements (which are of highly variable length) and in the specification of the nature of these elements. In this respect, however, the ASCII information separators have, at least by example, liberated the IBPS System from the Procrustean bed of the 80-column card and its fixed fields or from the tedious process of specifying the bounds of variable fields.

C. INPUT FUNCTIONS.

1. Input Media and Devices. Punched paper tape was chosen as the primary input medium for the IBPS System for the following reasons:

a. The fixed fields of punched cards were rejected because a *continuous record*, in which the elements remain simply of whatever length they are in the source bibliographies, was desired.

b. Although it is possible to obtain a continuous record on punched cards, an experimental use of cards in this way uncovered a serious problem in data correction: any error in any way involving *spacing* necessitates the repunching of all subsequent cards (often as many as eight or ten) in the sequence comprising that entry.

c. At the time the Pilot Study began, no keyboard equipment was available that could record directly, in computer-compatible form, onto magnetic tape.

d. The larger font (88 characters) afforded by a typewriter was considered to give it an advantage over the 47 character monocase of card machines.

The two major paper tape-punching typewriters are the Friden Flexowriter and the Dura Mach 10. The ultimate choice for the IBPS System was the Dura machine because:

a. Its punch mechanism employs positive retraction, rather than relying on springs.

b. Its tape reader is photoelectric.

c. Its readout speed is much greater than that of the Flexowriter.

d. It was hoped eventually to exploit the interchangeable-type feature of the Dura Mach 10 to print out polyglot bibliographical information--e.g., books and articles in Greek or Russian which some source bibliographies intersperse among others written in the Western languages.

N.B. Although the Dura Mach 10 is the *primary* input device in the IBPS System, provision has been made to accept, when desired, paper tapes produced by the Flexowriter or punched cards.

2. Data Base. One of the immediate objectives of the Pilot Study was to conduct a "trial bore" in the bibliographical corpus. The dimensions of this trial bore were prescribed initially by the following limitations:

a. Field within English Studies: American Literature.

b. Number of years to be handled: Publication in 1950-1963 inclusive.

c. Number of bibliographies to be processed: 14 bibliographies:

(1) "Annual Bibliography," *PMLA*.

(2) "Annual Bibliography of English Language and Literature" (MHRA).

(3) "Articles on American Literature Appearing in Current Periodicals," *American Literature*.

(4) "Folklore Bibliography," *Southern Folklore Quarterly*.

(5) *Literary History of the United States: Bibliography and Supplement*, 1962.

(6) *Eight American Authors, with Bibliographical Supplement*, 1963.

(7) *Dissertations in American Literature, 1891-1955, with Supplement*, 1956-1961.

(8) *The Year's Work in English Studies*.

(9) "Articles in American Studies," *American Quarterly*.

(10) "Bibliography," *Journal of American Folklore Supplement*.

(11) *The American Novel, 1789-1959*.

(12) *Poetry Explication*, rev. ed., 1962.

(13) *Abstracts of English Studies*.

(14) "Current Bibliography," *Twentieth Century Literature*.

Within the first year of research, however, it became evident that the writing and testing of computer programs must take first priority in the expenditure of time and energies, and that so considerable a data base was quite unnecessary for testing purposes. It was therefore determined that augmentation of the data base should be made a low priority task until the programming was essentially completed. At the time that this decision was made, the data base had assumed the proportions which it now has;

On magnetic tape:

MHRA	1959
PMLA	"
TCL	"
AL	"
DAL	1956-1961
Am Novel	1959 (articles) + all books.

On either paper tape or punched cards in varying degrees of readiness for transference to magnetic tape:

MHRA	1960-61
PMLA	"
TCL	"
AL	"
AES	1959
BAA	1955-1962
Poetry Expl.	1959 (articles) + all books.

3. Preparatory Editing. Four of the source bibliographies in the present data base contain only items on American Literature (or have sections which contain all such items). Five make no such distinction and so required editing. This editing consisted in

examining each item to determine whether it is relevant to American literature, language, or civilization. If relevant, the item was marked (a margin line) to indicate to the typist that it was to be transcribed. This editing function, in terms of the development of a complete Integrated Bibliography for English Studies, is gratuitous. It was made necessary by the Pilot Study's commitment to limit items in the trial bore.."

There are stylistic and subject-heading differences among various bibliographies. While it was believed from the outset that the computer could be programmed to recognize, utilize, and finally convert into unity the varieties of styles, as well as preserve those distinctions which aid retrieval, it was felt advisable in practice to provide a few additional cues, in case full expectations were not realized. In what admittedly was a mistaken conviction that a considerable data base would be necessary at an early stage of the project, a production operation for preparation of data for input to the computer was carried out prior to the completion of the programs whose success would guarantee that no such editorial functions need be carried out. Thus, *special* instructions were written for each bibliography, outlining a number of small variations from the general rule of "follow copy." Such instructions as "Precede each entry by Carriage Return and Tab," or "Underline all subject-headings and precede by Carriage Return and Tab" were provided. Such special instructions can, in fact, be dispensed with.

4. Paper Tape Production. So long as accuracy in the transfer of data from printed page to machine-readable medium depends on the

accuracy of a transcription made by a humanly-operated device (e.g., tape-punching typewriter or card punch), that accuracy will have to be checked; and while the computer can be programmed to recognize some kinds of gross errors (e.g., misspellings which produce non-words), it was no part of the present Pilot Study to explore or devise such programs. Therefore human proofreading was employed to check the accuracy of human typists and input devices. The operations performed are conventional and may be outlined briefly. First, however, it should be noted that, although most of the data punched was prepared by the Dura Mach 10 Typewriter, some data was punched on the Friden Flexowriter and the IBM 026 Key Punch.

The steps below, however, describe the production of paper tape on the Dura Mach 10.

a. Edited source bibliographies are transcribed, producing Tape I and hard copy #1.

b. Hard copy #1 is proofread for all inaccuracies in copy.

c. In the earlier stages of the project, Tape I was run through the tape-typewriter, producing hard copy #2 as a check on the accuracy of both the machine and the typist.

d. Hard copy #2 was then checked for obvious machine- and typist-produced errors.

e. Tape I is run through the tape-typewriter, producing Tape II, on which all errors noted in Step 2 or Step 4 are eliminated by stopping the automatic punching where errors occur and re-typing by hand; simultaneously, hard copy #3 is produced.

f. Tape II is read into the computer and stored on magnetic tape.

g. Largely for diagnostic purposes, stored data is, from time to time, read out via high-speed printer to check the operation of the input programs.

h. Magnetic-taped information is replicated on computer-produced Tape III, which is read out on the tape-typewriter, producing hard copy #4.

i. Hard copy #4 or the high-speed printout is proofed against the source bibliography, and all errors, deviations, and omissions are noted.

j. An appendix tape which will correct all errors, etc., noted in hard copy #4 will be punched and proofed like the other information. It is to be accompanied by the Override program, which will order the erasure of erroneous elements on the magnetic tape and their replacement with correct data.

D. TECHNICAL ASPECTS OF THE SYSTEM.

1. The Work to be Accomplished. The formation and utilization of an Integrated Bibliography implies the successful accomplishment, by computer processing, of a variety of tasks. These can best be considered as they occur in succession on the main line of production.

(a) Character Transformation. Characters (letters, numerals, punctuation, etc.) can be recorded in any of several media, and according to any of several codes.

Devices for direct keyboard generation of computer-compatible magnetic tape were not available at the time the system was planned. (Although some have subsequently come onto the market, they have remained very expensive; furthermore, a proofreading of the magnetic medium requires use of a computer.)

But the material transcribed by means of a keyboard into the alternative, a punched medium, is not in a form physically adapted to swift computer manipulation, and, after proofreading, will require a transformation into magnetic tape.

As to the code of bits used to represent characters, they vary from one device to another. (To name only three, there are the Hollerith punched card code and the Flexowriter and Dura Mach 10 punched paper tape codes.) It was therefore deemed desirable to translate all the native input codes into the new American Standard Code for Information Interchange as an additional preliminary to further computer manipulation of the data.

(b) Format Normalization. On a scale of a different magnitude, another standardizing process is necessary: the selected source bibliographies employ diverse typographical conventions to delimit the

several distinct elements which make up a Bibliogram, as well as to designate the position assigned to each Bibliogram in the Predicative (subject-classification) structure employed in each bibliography. In order that items derived from a plurality of sources may be automatically compared and merged, or searched selectively by element, it is necessary to interpret these diverse typographical discriminants into a uniform system of machine-sensible tags.

(c) *Integration of Bibliographies*. The character transformation and the format normalization having been achieved, the integration of the several sources can be undertaken. The several magnetically stored files are merged into a single file which can be sorted on any desired element with a view toward bringing into juxtaposition Bibliograms from different sources which probably refer to the same document. (For example, the succession "Sort on surname of Author" followed by "Sort on first n words of title--(perhaps omitting initial article)" would probably leave unjuxtaposed very few Bibliograms of identical referent.)

Upon completion of a merge sort the integration process is at a stage ready for comparing juxtaposed items, element by element, for determining, on rather simple principles, whether the differences between the items are compatible or incompatible with the possibility that the compared Bibliograms have identical referents. ("Compatibility" will in general mean that the only difference between two compared elements is one of the amplitude of the information contained; e.g., "Abel Baker Charles" is compatible with "Abel B. Charles" and "A. B. Charles"--perhaps even with "A. Charles" and "A. Baker Charles." "E. (for Ezra) B. Charles" would, of course, be incompatible.) Upon determination that two Bibliograms are co-referent, their logical union will be carried forward in place of either of them.

After the elimination of duplicates, the file can be re-sorted into some order more convenient for searching or more desireable for publication--such as an arrangement in a Predicative *structure* approximating one or more of those employed in the source bibliographies. (Of course, the Predicative *information* derived from all of the source bibliographies remains permanently available to each of the united Bibliograms. The element-specifying tags previously inserted are also retained.)

(d) *Retrieval.* The object of the retrieval phase of the system is twofold: the accomplishment of a property sort on the integrated file, and an arrayed printing out of the subfile answering to the property sort.

> The "property" upon which the sort is to be made--the specifications for conducting the search--are given as two elements: 1) the *Search Authority*, which specifies the character-string to be sought (phrase, word, or truncated forms of these); and 2) the *Rubric Authority*, which directs which types of location (elements of Bibliograms or ranks in the predicative matter) shall be searched.
>
> As an example, the Search Authority might be any one of the following:
>
> southern novelist(s)
> southern novel(s)
> southern
> south
> or
> southern novel....
> south....

and the Rubric Authority might seek the above character-strings within any or all of the Predicative headings and/or within any or all of the elements of the integrated Bibliograms (although for this particular Search Authority probably only "Title of Book" or "Title of Article" would be profitably specified).

Needless to say, more elaborately-structured search specifications can in principle, and ultimately will, be provided for: for example, specifications embracing *logical constraints* ("southern" but not "poem" nor "poet...," etc.) and those bringing *thesauri* into use ("Alabama," "Virginia...," "Confedera...," "ante-bellum," etc., on the one hand, and "verse," "lyric...," etc., on the other).

The "arrayed printout" phase of the retrieval process consists of eliminating the machine-sensible tags that had been inserted in the "normalize" process and replacing them with typographical discriminants like those utilized in printed bibliographies or LC catalog cards. Their exact nature varies with the kind of output and its intended use. At the most luxurious extreme, output intended for type-composition equipment can play with such resources as the utilization of bold- and light-face, roman and italic, lower-case-with-caps and small-caps-with-caps, and regular sized type with one or more reduced sizes--and all this selectively, so that on demand a format could be produced which would match that used in any of the printed source bibliographies, or in others. At the most modest end of the scale (to skip over outputs of intermediate quality such as microfilmed cathode-ray-trace and various kinds of typescript) there is available output from the crude old dependable line-printer, with its very limited monocase font. For this, an intelligible arraying involves much reliance on an elaborate system of vertical spacing, horizontal indentation and perhaps some fixed fields. In any case, the arraying process seeks to restore automatically to human readability that which had in the normalizing process been made machine-sensible.

(e) *Oblique functions.* The foregoing processes form one main line, from input to output. There are at least two functions which stand outside that main line but which must nevertheless be regarded as part of the total system: Correction, and Authority Compilation.

(e.i) *Correction.* No matter how great the pains taken, by way of proofreading and re-keyboarding, to produce flawless input files, some errors will inevitably crop up in the proof output of the integrated file; means of correcting it without returning to the beginning of the whole process of compilation must be provided. "Updating" (addition of individual new items) and "Expansion" (incorporation of entire new bibliographies) are processes closely allied to correction; however, it will easily be recognized that these scarcely differ from the "integration" process described above, while correction implies also the "property sort" capability of the retrieval process--if only to locate the items to be corrected. Nothing essentially novel or tricky is required for any of the three processes.

(e.ii) *Authority Compilation.* The larger part of the material

with which this Pilot Study is concerned consists of journal article citations, and in these the journal titles are almost invariably abbreviated. The fact that there are discrepancies between the abbreviation systems employed in the several sources imposes an obstacle to the merging process, an obstacle which is best overcome by compiling tables interpreting the abbreviations used in each source ("Received Journal Authorities") and, within each bibliographical file, automatically replacing the original abbreviations with their corresponding full titles in preparation for the merger. (Subsequently it may be advantageous to merge the several received authorities themselves--here supplementing automated merging with considerable manual and intellectual work--into one "Canonical Journal Authority" for making a uniform reduction, in the integrated file, to an IBPS-standardized set of abbreviations.)

Other "authorities" will in future be useful: e.g., Publisher, Personal Name (of Author or of Subject), University (or other corporate) Name. The process of their compilation will proceed analogously, except that less human work may be required in view of the information that may be garnered from within the united Bibliograms. (E.g., one of the sources may chance to have given the name of an Author, Abel Baker Charles, in full.)

2. Segmentation of the System. In the foregoing section, the system designed for the Pilot Study has been described as though segmented into six programs:

(a) Character transformation

(b) Format normalization

(c) Integration

(d) Retrieval

(e.i) Correction

(e.ii) Authority compilation

As a matter of practice in its construction, however, the system has been segmented in a much more detailed fashion. Several considerations underlay this approach.

(a) Technical. It was scarcely to be expected that the merge sort processing could be conducted using traditional string-matching techniques, in view of the commitment to the "follow copy" rule and what that implies in the complication of the record caused by the occurrence of such things as diacritics, foreign alphabets, and varieties of typographical style even in our domestic alphabet. Many such details would be "non-substantive" with respect to the property-sort portion of the retrieval phase (but would regain substantivity at the "array" portion, in certain cases). It seemed simpler to anticipate the necessity of making the property sort on a (substantive) character-by-character basis--a not unreasonable decision in view of the increasingly higher speeds of the modern computer. Consequently, the large word-size of the larger and more expensive computers would not be efficiently exploited, nor would the available techniques of the compiler languages commonly used with them.

(b) Strategy considerations. As can be appreciated, the total system is one of some complexity. Rather than attempt to program--

and then to debug--the system as a whole before receiving some assurance of the soundness of the work--it was decided to seek, at each stage, assurance in the form of visible, tangible output--even if it were partially-processed, interim output. The task of realizing the system may be likened to that of throwing a bridge across a broad chasm--from input to ultimate output--and the soundness of each span must be tested before trusting it as a point of attachment for the next.

(Of course, underlying the above rather subjectively-expressed point is the well-known fact that the difficulty of debugging a totally unproven program tends to increase, not directly as the number of its component instructions, but rather as some power of that number.)

(c) *Economic considerations*. A commitment to a small computer and its associated assembly language, while entailing greater tedium in program encodement, promised the advantages of greater control over the economical design of our programs and the simplification of the debugging process. But the outstanding reason for the choice of the Control Data Corporation 160-A Computer (ideal in view of its integral paper tape reader) was that access to such a machine, during its idle moments, was offered to the IBPS--not only free of charge but on an "open shop" basis. Because of the amount of debugging needed for a system design which was in so many details a pioneer, the offer was one which hardly could be refused--debugging through a closed shop is measured, not in minutes per bug, but often in days or weeks per bug.

As a consequence, the system was segmented into a rather large number of programs almost all of which could be handled by a computer no more imposing than the 160-A, and several programs were designed to produce interim outputs for diagnostic tests all the way across the "bridge." (The first two of the diagnostic programs, MPAT and MTAN, consisted in large part of logical modules which could, as a matter of sensible programming economy, be reincorporated in later programs, of both the diagnostic and production types.

It was never expected that the "merge sort" and "optimize filing sequence" elements of the "integration" span of the bridge could operate on a computer as small as the one which sufficed for the rest of the main line; these will need to be designed for a machine with a larger core. But a working model of the last span of all, the property sort and output array, has demonstrated those processes to be well within the capabilities of the little computer.

Of course, upon the final checking-out of all the many elements of the total system, it will become advantageous to consolidate them, first into the six functional blocks listed above, and eventually into an integral but compiler-addressable system.

3. Evolution of Programs. The construction of each segment of the system--each program--has followed, in general, a common succession of stages:

(a) Preparation of the Logical Diagram ("flow diagram"). The purpose of the program is to be achieved in a structure of connected sequences of logical operations upon the contents of a data file of some kind. It is this structure that is displayed in the logical diagram. The diagrams are *general* (in the mathematical sense): they display only the embodiment in programs of the required logical

manipulations and their interconnections; they do not contain, for example, computer instructions specific to any given piece or make of equipment, nor "statements" peculiar to any programming language they are therefore adaptable with substantially impartial universality to whatever programming language a programmer may desire to use. Great pains have been taken to ensure that the symbolism and notation employed in these diagrams shall make them intelligible to future programmers (see "Legend to Programs," Appendix A). So far as practicable, their interpretation has been facilitated in other ways, notably by

(i) The vertical or horizontal alignment of comparably significant operations (notably the indexing operations).

(ii) The unmistakable designation--an outflow line at *right angles* to the inflow line--given, wherever the line of logic bifurcates, to indicate which outflow line is to be regarded as "normal" sequence and which as a conditional jump from that norm.

(b) Manuscript encodement. On the basis of the general logic exhibited in the flow diagram the implementation of that logic for a specific computer is made in the phase of programming called "coding."

The IBPS programs thus far have been coded in an assembly-system language developed as factory software for the 160-A computer by its manufacturer, Control Data Corporation. Named OSAS (for One-Sixty Assembly System), it is a language relatively unsophisticated compared to compiler languages like COBOL or FORTRAN, commonly used on larger equipment; however, the programs generated by it are far more economical in their use of computer core space and the time required for their execution. In fact it differs from "pure" machine language only in such respects as are helpful to the person doing the encodement. These are principally the following:

(i) Mnemonic rather than numerical instructions are used (e.g., for "subtract the constant" SBC rather than...)

(ii) Locations can be assigned symbolic names rather than having to be specified by accurately counted numerical addresses.

(iii) Programming blunders of certain kinds are caught at assembly and thrown back in the programmer's teeth.

The factory-supplied OSAS Coding Form has been found unsatisfactory for this work. There is room on the form only for the information required by the card-punch operator; but she does not need to know as much as the programmers need to know and keep track of. Consequently a form is used which has four columns added at the left of the leftmost column. (See "Legend to Programs," Appendix A). In these added columns notes are kept on (1) the settings of the four storage bank controls, (2) changes in the contents of the arithmetic register, (3) changes in the contents of storage cells, and (4), at places where two lines of logic rejoin, a record of the source(s) other than the one directly antecedent to the join-point.

The 11" x 17" sheets making up the manuscript encodement of the "operational" (non-tabular) part of the program are, for convenience' sake, Scotch-taped into a continuous scroll. While the coding is being carried on, numerous insertions and deletions inevitably prove necessary, and the scroll format, by means of cuts and splices, maintains the sequence of the program more reliably than would a sheaf of loose papers.

Tab cards: Only then are the pertinent contents of each line of the scroll punched into a tab card. The deck of punched cards, which is by paper tape standards a very heavy and bulky way of storing any kind of information, nevertheless lends itself to convenient further correction--which invariably proves necessary, either upon initial assembly or upon trial operation with specimens of data.

(c) Binary paper tape. The 160-A Computer is then loaded with a program called "OSAS A/H 7130 BiOctal Assembly Program," whereupon the program deck is read by a card reader (CDC 405). The assembly process then proceeds to yield two products: (1) A high-speed printout of the assembled program in machine language, and (2) a punched paper tape embodying the program. If the program is free from errors (never the case at the first attempt!), this tape then supersedes the deck of cards. It is far more compact, it is lighter (by a factor of more than 150), and it can be read much more quickly.

(d) BiOctal paper tape. The paper tape produced by the assembly program is not yet, however, in the form most convenient for repeated use. It is a "binary" tape, and each time it is to be

loaded into the computer it requires the assistance of a program called "AA4.01 Binary Paper Tape Loader." However, it need be so loaded only for one first time, whereupon it may be translated into a form not requiring such assistance--a form that can be loaded directly. This translation is accomplished by use of a program called "TRK's Relocatable (40_8) BiOctal Dump." The product is another, and definitive, "BiOctal" paper tape. The binary tape is then discarded.

(c) *Magnetic tape*. Ultimately, all the programs composing the system are to be stored on magnetic tape, replacing the present collection of fractional-ounce paper tape coils.

4. Description of Programs.

Common features. The significant function of most programs is carried out by serially processing *paragraphs* of the text--the paragraph being treated as the information unit next larger than the individual character. The space-separated words of the text, or lines of the typescript are not, in general, recognized as significant. On the other hand, external storage of data is by *blocks* of a constant number of characters. The source material is *read* a block at a time and, if output onto magnetic tape, *written* a block at a time. Requirements for core storage space to hold one block of input data, one block of output data, and a paragraph's-worth of characters undergoing processing internally, must all compete within the computer amongst one another and against the space requirements of the program itself. The standard block size has therefore been fixed at approximately 2000 characters--slightly less than one quarter of the total core of the 160-A.

(a) PTMT (Paper tape to magnetic tape), logic and coding by Curtis D. Benster and Eugene P. Schumacher, is the initial segment of the system. It converts input paper tape data to the magnetic tape image without any code transformation or other alteration except the omission of deleted frames of the paper tape and of "pseudo carriage returns" mentioned below. It is intended to transfer the data from a rather fragile medium to a more dependable one--which is, more importantly, a processable medium.

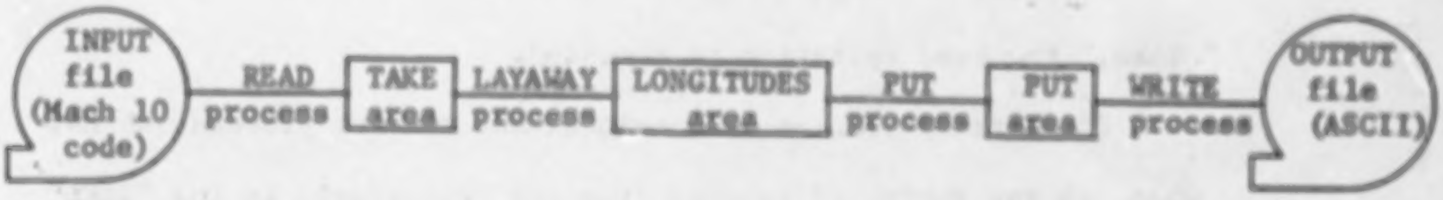
(b) MPAT (Magnetic to paper tape), logic and coding by Curtis D. Benster and Randy Levine, is primarily a diagnostic program, inverse to PTMT above, intended to test it, but its tape output is capable also of being used for making corrections manually when necessary. To facilitate the latter purpose, the output of MPAT is modified in two respects (neither of which substantively alters the data in any way): the paper tape is physically marked off into special blocks of sixteen paragraphs, each separated from its neighbors by a few inches of leader; to generate a comparable spacing between 16-paragraph blocks in proof typescript, "pseudo carriage return" characters are supplied.

(c) MTAN (Mach typewriter code to Anelex line printer code), logic and coding by Curtis D. Benster; revised by Curtis D. Benster and Randy Levine. A diagnostic program adapted to detecting garbled punching by the Mach 10 typewriter. The "Write" module of this program is also used in other programs having high-speed line printing as output. The printed lines correspond generally to those of the typescript, although provision has been made to accommodate situations in which the typewriter's "carriage return" character has

been mangled or lost. The printout is in arbitrary blocks of sixteen paragraphs.

(d) MASC (Mach code to American Standard Code), logic and coding by Curtis D. Benster and William S. Crosby, is a magnetic-tape-to-magnetic-tape code translation program receiving data in the Dura Mach 10 typewriter code and forwarding the data in the American Standard Code. It will be described in some detail, since it illustrates features common to a number of our programs.

File specifications: Non-packed (one 7-bit character occupying two consecutive frames on half inch magnetic tape). The last record proper of the file concludes with the code sequence "carriage-return"-"tab"-"end-of-record," and is succeeded by the quasi-record consisting only of the "end-of-file" code. In the ASCII output file the highest-order (lefthandmost) bit of each "word" containing a character regarded as "non-substantive" is set equal to 1. ("Non-substantives" include shift-codes of various kinds and other material whose consideration should not be permitted to complicate or delay early versions of the Pilot Study's search programs.)



Read operation: A block of data is Read from the input tape and written into the "Take" area of core storage. This is then a bit-for-bit replica of the input tape, which was a virtual replica of the source paper tape (i.e., the magnetic tape differed only in having been purged of stops and deleteds.)

Ordinarily the Read operation concludes when the "Take" area is filled to capacity--(this usually coincides with the Reading of an "end-of-record" code on the magnetic tape); but if the record being Read is a short one, termination of the Read operation is effected by an encounter with that code regardless of the extent to which the "Take" area has been filled.

When the Read process eventually encounters the "end-of-file" code on the input magnetic tape, that reel of tape is rewound.

Layaway operation: One at a time, the characters stored in the "Take" area in Dura Mach 10 code are read, and, in most cases, referred to a Code Translation Dictionary (called MO) which furnishes the equivalent American Standard Code--with some supplementary annotations--and this dictionary form (annotated ASCII) is then Laid Away in the "Longitudes" area of core storage. For those characters that present special problems, annotated "pseudo-ASCII" forms are furnished by the dictionary, and they are given the necessary special treatment either in the course of the Layaway or the Put operation.

Whenever the Layaway process exhausts the material stored in " Take," the Read operation is repeated.

The Layaway process is concluded and the Put process begins, when, in the course of reading from the text stored in the "Take" area, the code sequence "carriage return"-"tab" is read. [This amounts to saying that only one paragraph at a time is Laid Away. In the unlikely event of a single paragraph's overflowing the maximum capacity available for use as "Longitudes" area (several thousand characters), the computer will halt at the point in the program called "CRAMPT."]

The Longitudes Area is a reconstructed replica of the typescript, in that it is indexed forward, by one cell upon the Laying Away of each code standing for any graphic character or space or carriage-return, and indexed backward by one cell upon receipt of a back-space code; the index remains stationary upon receiving an up- or down-shift code.

Processing in Longitudes Area: Each character that is laid away in the "Longitudes" area is provided with annotations. These are destined to be expunged before output, but are useful internally at various stages in the operation of the program. The annotations come from two sources: (1) Some annotations are supplied from context, (2) Others are furnished by the Transformation Dictionary, and state certain inherent properties of the character.

In each 12-bit cell of the 160-A core storage, the ASCII code occupies the low-order (righthandmost) seven bits of the cell, and its annotations occupy the remaining five bits. In the notation being used, the bits that spell out the code are numbered—reading to the right—from b_6 to b_0 inclusive. The bits remaining at the left, notated in a scheme whose mnemonic is "Run, alphabeta," complete the computer word, thus:

Computer "Word"	
Annotation	Character
R u n α β	b_6 b_5 b_4 b_3 b_2 b_1 b_0

The R-bit is supplied by the Transformation Dictionary, and, if it equals 1, indicates that this character, coded as a pseudo-ASCII, requires special handling by a subroutine. The character will be referred to the point in the program called "RATS" (Random Access to Subroutines). Two examples of such characters: Lower-shift 1, which must be diagnosed as either a numeral "one" or a letter "ell"; and the slash (/), which might be the boundary of one of the established IBPS Escape Symbols, or might merely be a slash itself, as in "and/or."

The value of the u-bit is supplied from context; it is set equal to 1 if the longitude is ultimately found to be *underlined*.

The n-bit is furnished by the Dictionary, and equals 1 if the character is one of the numerals.

The alpha-bit given in the Dictionary equals 1 if the character is a letter of the alphabet.

The beta-bit has either one of two functions:

(1) If the character is a letter of the alphabet, then it is ascertained from context (most recent shift-code) whether the letter in the typescript was capital or lower case. If the former ("Big"), ? the beta-bit is set equal to 1.

(2) For the space, carriage return, period, comma, and hyphen, the Dictionary provides a beta-bit equal to 0; for all other non-alphabetic graphics (including all numerals) the Dictionary gives it as 1. The peculiar property common to the above-listed five characters is that this program will regard an italicized (underlined) passage as continuously italic even where the typist has not given it a continuous underline, provided that the non-under-

lined gaps contain only these five characters or combinations of them. All other characters are Barriers to the gratuitous extension of the underline, and, therefore, if their alpha-bits are equal to zero, their beta-bits, furnished by the Dictionary, equal 1.

If the typewriter carriage was backspaced in the course of typing a line, and a second character struck in a longitude cell already occupied by another character, this program will regard the most recent character as obliterating and replacing any earlier occupant of that longitude cell. [Exceptions: (1) Of course, an underline will not obliterate anything, but will only affect the u-bit in that logitude; (2) a space will not obliterate anything; and (3) nothing has power to obliterate an underline.]

The occurrence in text of a *slash* (/), followed, at two, three. or four removes, by another slash, will initiate a search, in another Dictionary, one comprising the IBPS Escape Symbols (DICTES), to determine whether the characters enclosed between the slashes constitute the kernel of such a symbol. If they do, this first slash is the symbol's left-hand boundary; the right-hand boundary of the symbol is then transformed to an ASCII "reverse slant" (counter-slash), and both boundaries are appropriately annotated. [Annotated as to being boundaries of ASCII-realizable symbols, for example, and/or bounding "alphabetics" (e.g.: /ae\), "shifters" (e.g.: /bf\ = bold face), etc.]

The occurrence in text of the ambiguous lower-shift character l initiates a scan of its context to determine whether it is to be rendered as the numeral "one" or as the letter "ell." Typists can be instructed how an indelible numeral "one" or an indelible letter "ell" can be written at the keyboard to override the AMBIGU subroutine in the rare cases when it would make an incorrect diagnosis. (She will type the IBPS Escape Symbol /one/ or /ell/.) However, a need for this has not yet been encountered in the IBPS data--all 1's occurring in it have thus far been diagnosed without error.

After processing in the "Longitudes" area has been completed, and as the Put operation is about to begin, the characters stored in "Longitudes" stand, in general, encoded in ASCII supplemented with complete annotations.

Those characters that were realizable on the Dura Mach 10 typewriter, but that are absent from the ASCII set (namely, ¢ and °), are left in a pseudo-ASCII form, but are now tagged for later expansion into a program-constructed IBPS escape symbol (i.e., /ct\ and /dg\).

Those characters that are realizable in ASCII as single characters, but which were absent from the Dura Mach 10 set (such as the mathematical sign for "greater than"), would have been ordered by the typing of an IBPS Escape Symbol (e.g., /gt/). Such symbols remain at this stage in their expanded form, but they are now tagged for later realization.

Put: The purpose of this program is to set up, one record at a time, the data that are to be replicated onto the output magnetic tape. The characters are read one at a time from the "Longitudes" area, and, in general, are purged of their annotations, and written directly, character for character, into the "Put" area of core storage. But those characters tagged for expansion and those character-groups tagged for realization are now transformed, and written into the "Put" area in their expanded or realized forms respectively.

As the reading in the "Longitudes" area progresses, the occurrence of a change in the non-underlined/underlined status of consecutive characters orders the writing of an /it\ (Italic shifter) or $ (counter-shifter) Escape Symbol, before the character itself is written--this, subject to the exception that the occurrence of non-underlined instances of any of the five "non-barrier" characters (space, carriage return, period, comma, or hyphen, or of combinations of them) does not order the writing of a counter-shifter unless the next "barrier" character is non-underlined.

The occurrence of a change in the capital/lower-case status of the alphabetics as they are encountered (not necessarily in *immediate* succession; non-alphabetic characters are irrelevant to this) orders the writing of the ASCII "shift-out" or "shift-in" code before the alphabetic is written. ("Out" to capitals; back "In" to the normal: lower case.)

Whenever the Put process exhausts the material stored in the "Longitudes" area, the Layaway operation is repeated.

Whenever the "Put" area is filled to capacity or when, ultimately, an end-of-file code is read, the Write process begins.

Write: The contents of the "Put" area are now replicated, character for character, onto the output magnetic tape and followed by an end-of-record code and an inter-record gap (0.7 to 0.9 inch of blank tape). After the Writing of each block of data but the last one, the Put process is repeated. After the last block of data has been written onto magnetic tape, an end-of-file code is written and the output magnetic tape is rewound.

(e) ASTAN (American Standard Code to Analex code), an adaptation by William S. Crosby of MTAN (Mach to Analex) by Curtis D. Benster, reads data recorded on magnetic tape in ASCII, prints it on high speed line printer optionally with or without indications of changes in shift status. (A printout with such indications provided is essential for proofreading; one which suppresses them is less cluttered and more legible; it would be preferred by most users.)

CNFORM, logic and coding by Curtis D. Benster and William S. Crosby, is a prototype of the format-normalizing programs--those whose purpose is to adapt unedited bibliographical source material to the process of automatic searching and automatic merging.

Toward this end the program identifies the several "elements" of each bibliographic item (e.g., "Subject," "Author," "Title") and applies to each element a distinctive machine-sensible tag.

The traditional approach to such problems of separating elements has been to impose upon the keyboard operator the editorial responsibility of marking their boundaries, very commonly by utilizing fields assigned for the purpose within the 80-column tab card. The IBPS rejected this conventional approach for two reasons:

(i) Tab card technology seemed needlessly confining with respect both to the 80-column limitation and to the meager font of characters provided.

(ii) More important was the fact that this Pilot Study is expected to explore the potential of automation to the utmost, and a part of that potential which is quite within the foreseeable future involves the supersession of visual-manual transcription methods by the use of optical scanning devices. Since such devices will have no inherent editorial capabilities at all, the IBPS System should even at this stage attempt to incorporate in its programs the capability of automatic segmentation of bibliographic entries. Segmentation into elements is accomplished by interpreting those typographical discriminants, furnished in the printed source bibliography and literally transcribed in typescript, which are normally utilized by a human reader of a bibliography to make the same segmentation mentally.

The first of the segmentations performed by the CNFORM program is between Bibliograms and predicative matter. This segmentation takes account of the fact that predicative matter may sometimes be found deposited in the interior of a Bibliogram. (E.g., there may be an annotation supplied by the bibliographer to explicate a title whose meaning is unclear.)

The two identified kinds of material are then subjected to processing directed respectively to two different ends: predicative matter is usually organized hierarchically, and the rank and position of a given Predicative must be specified; most of the information within a Bibliogram, on the other hand, is of one rank--Author, Title, Imprint, for example, differ from one another *qualitatively*, not *quantitatively*--but to the greatest extent possible the Bibliogram must itself be segmented and the nature of each qualitatively different segment clearly specified.

Nesting. There is one important qualification to be made to the assertion that all segments of a Bibliogram are of equal rank: this has to do with those situations in which documents are "nested" within documents--the most common being the case in which an article is nested within a journal. Each of these qualifies as a document, e.g., each has a title; the article will have an author; the journal will usually have a person, namely its editor, standing in the relationship to it of an author. But this is only the simplest case of what may exhibit at times a structure of great complexity. A book, for instance, entitled to a Bibliogram in its own right, may be referred to only obliquely in a bibliography which mentions articles reviewing the book, each such review article having perhaps an author and a title of its own and appearing nested within a separate journal. And this is far from exhausting the possibilities of complication.

Since it is not possible to anticipate the detailed structure of all possible complicatedly nested Bibliograms and to set up a notation capable of tagging each specific location in each specific type of structure, an approach to the problem is required which, by the employment of a few principles of universal capability, will enable all conceivable structures to be represented. (An analogy: to represent the whole numbers up to, say, ten million, we do not need ten million distinct characters; we use only ten, combining them as needed according to a simple rule; and as a result we find ourselves not restricted to ten million as an upper limit.)

Accordingly, an approach has been adopted which declares that all Bibliograms are essentially tripartite--they consist of the three segments Author, Title, and "Other Material"--and any specific bibliographical entry of one of the complex types mentioned above should be dissected into two or more mutually intersecting Bibliograms (neither of which need be fully complete) and that the relationship between them should be specified as a situation in which one is "nested" within the other. Thus the journal article's Bibliogram would be segmented as follows: Author, Title (of the article), Other--and, within "Other," a tag indicating that what follows next is part of a Bibliogram of a different rank (namely, one relating to the Nest document--the journal within which the article was nested); this Nest Bibliogram will usually be incomplete in that there will not commonly be mention of Author (i.e., Editor); but Title (of journal) will be given. Pagination is regarded by us as a property of the given article rather than of the nest journal, so that after giving the title of the journal the compound Bibliogram will include an "out-of-nest" tag and revert to the Other element of the original Bibliogram. By similar means it is expected that all complicated bibliographical items can be reduced to a nested structure of simple, tripartite Bibliograms.

(i) Operation of the Program.

Input File. In its present state, the CNFORM program accepts from magnetic tape a record representing, in characters of the American Standard Code for Information Interchange (X3.4-1963), a literal transcript as made by a typist, of bibliographical material as given in the format of the annual bibliographies of *Publications of the Modern Language Association of America*. (It is expected that the modifications needed to accommodate bibliographies in other formats can be made without difficulty, since the program consists essentially of linkages between a small number of subroutines.) The processing to which the input file has been subjected has supplied "non-substantive" tags to characters representing shifts, in order to expedite searches.

Output file. At this stage in the evolution of the program, a complete tagging of predicative matter has been accomplished, and a sufficiently detailed degree of segmentation of the Bibliogram has been achieved to demonstrate the practicability of automatic segmentation.

Specific tags are applied to Author, to Title-of-Article as distinguished from Title-of-Book; Title-of-Nest as distinguished from both. Since the explicit tagging of titles makes redundant the italicizations or quotation marks by which they had been implicitly tagged in the original record, these typographical features are at this point expunged. Similarly, but with respect to predicative matter, the explicit tagging of cross references makes redundant such expressions as "See" or "See also" which appeared in the original, and these are now dropped from the record. The square brackets that had served to delimit annotations supplied by the bibliographer are replaced by a predicative tag of appropriate rank. All other predicatives are assigned the tags proper to their places in the hierarchy.

Finally, production details appearing in the input file (sign-ons and sign-offs by typists) are expunged.

Thus the output file is in a condition enabling it to be searched selectively by imposing on the search the requirement that the Search Authority (the word or phrase which is to be the object of the search) shall be compared only with predicative headings, for example, or only with matter tagged as Title of Article. Thus searches will be expedited, and in the future the segmented and tagged file will lend itself to the operation of merging.

In broad outline, the CNFORM program resembles the program called MASC which has been described in detail elsewhere. Like MASC, CNFORM utilizes three areas in the computer's core storage for data. Of these, one, called the "Take Area," accepts data in blocks of about 2000 characters as read from the Input File; another (called the "Put Area") assembles the processed matter ready to be written, one block at a time, on the output tape. The central area, called "Layaway," is a place in which one paragraph at a time is marshalled and processed.

(ii) Further Development. At the present time CNFORM is working well, within its limitations. The areas in which further developmental effort can most profitably be expended, in the order of descending priority, are:

(1) Within the Author field, the multiple-author problem has yet to be tackled.

(2) The nesting concept has yet to be implemented with respect to review articles.

(3) Further segmentation of the Other field should be undertaken. (E.g., Volume, Issue, Date and Pagination of journal article can in principle be individually tagged.)

(4) A third kind of cross-reference (one given in the form of a bibliographer's annotation) needs to be distinguished and tagged.

(5) "Italics within italics" (e.g., book-title nested within book-title) are in *PMLA* rendered by a redescent to roman type rather than by an ascent to small caps. This situation needs to be accommodated in the logic of the program.

(6) "Quotes within quotes" (e.g., poem title nested within article title) are, by the usual conventions of American typography, printed as single quotes. When the nest title has been tagged as such and its original double quotes are discarded as redundant, the inner single quotes should be promoted to double quotes. This is not achieved by the program in its present state. An apostrophe-diagnosing subroutine will be needed for the accomplishment of such a promotion. (Eventually it may be desireable to replace internal quotes by a tag meaning "Presumptive title--of poem or article.")

* * *

This ends the account of the programs of the system completed for the Integrated Bibliography Pilot Study to date. These several programs have been thoroughly checked out, on the very elaborate and varied data of the Integrated Bibliography Pilot Study.

Although they are correctly regarded as complete and ready for use in production, modifications will have to be made in them in accordance with the evolution of the American Standard Code for Information Interchange.

E. RETRIEVAL.

1. Evaluation of Capabilities. The retrieval capability of the IBPS System appears to be great, but it has not yet been tested. An investigation into the indexing problem of the IBPS, conducted by Jessica Melton of the Center for Documentation and Communication Research, Western Reserve University, found that, although titles of studies in the field of American Literature are not, in themselves, sufficient to indicate adequately the content of the studies, the titles plus certain predicative matter, i.e., the subject-headings and annotations--and abstracts when they exist--are a great deal more informative.[1] An accurate assessment of the value of predicative matter for retrieval cannot be made, however, until more evidence is available, i.e., until several bibliographies have been merged and the sum of the predicative matter from the various bibliographies (as it is attached to the merged [final] entries in the Master File) has been examined. Adherence to the "follow copy" rule for data at input precluded the possibility of employing any existing indexing and retrieval system which depends upon the assignment of indexing tags prior to input. This is perhaps just as well, for nearly all existing retrieval schemes are conversions from mechanical (e.g., the card-sorting techniques of the Uniterm system) or manual (e.g., Keydex) retrieval systems and depend heavily on human indexing. On the other hand, automatic indexing systems (insofar as they exist) appear to be somewhat less reliable than the proposed IBPS methods and are de-

---

[1] For a complete summary of this investigation, see *The IBPS Progress Report*, pp. 73-75.

pendent upon scanning entire texts.

Much testing must take place before optimum use can be made of the IBPS retrieval capability, but it is possible at this stage to describe the capability and the uses to which it will be put. The limitations of the present project have dictated that it employ, for retrieval, word indexing, in its broadest sense, as opposed to subject indexing.[2] That is, every word which appears in the data base--whether subject-heading, author, title, annotation, abstract, or journal volume number--is, without change, a potential indexing term. The usefulness of this information is enhanced by the following facts: (1) as a result of "following copy," no predicative matter (subject-headings, annotations, abstracts) has been lost in transfer from printed page to magnetic tape or in manipulation within the computer; (2) the IBPS program CNFORM identifies and labels, automatically, all predicative matter, as well as the traditional segments of the bibliographic citations (author, title, publication data, etc.); and (3) the connection between the bibliographic items and the predicative matter with which they are associated, in all the bibliographies in which they appear, is not lost, but is available in a merged Master File for searching purposes. For example, an item might appear as follows in two source bibliographies:

---

[2] "Word indexing" describes a scheme whereby words are taken directly from the source, without any change, for use as indexing terms. In "subject indexing" (sometimes known as "concept indexing") an indexer determines which words are to be used as index terms, regardless of whether they appear in the source.

Source #1	Source #2
I. American Literature	I. Fiction
A. ---	A. ---
B. ---	B. ---
C. ---	C. Authors
D. 20th-century.	---
---	---
---	---
---	*Hemingway*
Hemingway.	---
---	Jones, John Howard, "Hemingway's Heroes." *MFS*, X (Fall 1959), 12-49. [Real-life models of Frederick Henry and Nick Adams are identified.]

Jones, John H. "Hemingway's Heroes," *MFS*, X (1959), 12-49.	

Such an item will be ordered in the merged Master File so that the information is brought together as, in effect, one entry:

Jones, John Howard, "Hemingway's Heroes," *MFS*, X (Fall 1959), 12-49.

Source #1
I. American Literature
D. 20th-century
Hemingway

Source #2
I. Fiction
C. Authors
Hemingway
Real-life models of Frederick Henry and Nick Adams are identified.

Although nothing in the IBPS System is incompatible with a future incorporation into the Master File of researched information in the form of subject index terms (it can be so incorporated simply by treating the additional terms as further predicative matter), such a plan would be most expensive and time-consuming for a full-scale Integrated Bibliography; it would mean the examination of hundreds of thousands of books and articles for the purpose of indexing, or re-indexing, by a trained staff at an estimated minimal rate of 12.5 man years per 100,000 items. This was one of the major reasons which determined that in this Pilot Study there must be no looking behind the printed source bibliographies to the original documents--books, articles, dissertations, etc. This does not preclude, however, the possibility of employing a staff to examine printouts of a merged file (which includes, as illustrated above, the totality of the predicative material associated with each item in every bibliography in which it has appeared) and to assign indexing terms from a standardized list on the basis of titles and predicative matter. But since this too is a time-consuming task (approximately four man years per 100,000 entries), it is felt that such a procedure should not be undertaken unless the more easily attainable automatic word indexing method proves itself a great deal less effective than it presently promises (though the pitfalls of its relatively uncontrolled vocabulary are known).

These necessities and considerations have determined that the IBPS employ for its first retrieval experiments the more easily attainable system which is, in effect, a term-on-item word index with a vocabulary controlled by a thesaurus. The IBPS search routine, now operative, permits the search for the occurrence of any word or words anywhere in the data base. The affixing of information separators to all segments of the bibliographic entries, including the predicative matter, makes possible, in addition to searches on every word in the entire merged unit record, searches which are limited to any of the following discrete segments of information concerning individual bibliographic entries:

1. Author, editor, or compiler of item, e.g., Kazin, Alfred; Leary, Lewis.
2. Publisher, e.g., Books published by Arkham Press.*
3. Journal, e.g., Articles in *Encounter*.
4. Journal volume, e.g., Articles in vol. IX of any magazine(s).*
5. Year, e.g., Items published in 1959.*
6. Books, i.e., Items not appearing in journals, *Festschriften*, or other collected editions.
7. Articles, i.e., Items not published separately.
8. Keywords in titles, annotations, and abstracts, e.g., Naturalism, O'Neill, *Sartoris*, poems, poetry.
9. Subject-headings of source bibliographies, e.g., *Semantics*, *Satire*, *Criticism*, *Drama*, *Myth*, *Colonial*, etc.
10. Surname subject-headings, e.g., *Cable*, *Clemens*, *Eggleston*, etc.[3]

One through *seven* do not offer the kinds of searches that literary scholars generally require, but the availability of these searches should open up a number of new possibilities for studies of recent scholarship and criticism which were not formerly feasible because of the difficulty of gathering data.

Eight, *nine*, and *ten* will be the more important and more frequently used search areas. It is from these areas that the computer will be able to identify (in a sense in much the same way the researcher reading through a bibliography identifies) the occurrence and co-occurrence of words which indicate the content of the entry; but a computer search is more comprehensive and, of course, a great deal more rapid than any search by a human. In compiling a bibliography on, say, Allen Tate, the researcher would turn in the *PMLA* Bibliography for 1958 (only one of the many bibliographies he must laboriously consult) to the surname subject-heading *Tate*, where he would find two articles and one cross-reference to works on Tate. Seldom, if ever, would he plough through the 20th-century *Miscellaneous* section; but unless he did, he would miss at least one article dealing with Tate and several dealing with the Vanderbilt Agrarians in general. A computer search would not miss these items, for it would search the entire file. To indicate further the greater comprehensiveness of a computer search of the sort intended, the same *PMLA* Bibliography for 1958 has entries in the 20th-century *Miscellaneous* section about, among others, Donald Hall, Herbert Gold, Norman Mailer, and William Styron with no cross-references to them from their appropriate surname subject-headings; this list of items which the IBPS retrieval system will be able to retrieve but which all but the most painstaking researcher would miss could be extended almost *ad infinitum*.

---

*Not yet identified as such by information separators, but such identification has been provided for and is planned.

[3] Surname subject-headings are not actually different in kind from subject-headings (*nine*), but they are of such importance that they are here listed separately.

The same claim for greater machine efficacy possibly does not hold in searches for the co-occurrence of terms, but the similarity of man and machine methods remains, and it is hoped that future studies will show that the results are much the same. A seeker for items on the style of Henry James would, for instance, find one item in the 1958 *PMLA* Bibliography which is unequivocally on that subject: Krause, Sydney J. "James's Revisions of the Style of *The*

Portrait of A Lady." A computer search would retrieve the same item, and both the man and the machine would select that item because of the co-occurrence of the words *James* and *style*.

Any word or words from any of the ten search areas above may be requested singly or in any combination of logical sum, product, or complement.

LOGICAL SUM represents an *and/or* function. That is to say, a search for the logical sum of A and B would produce all entries which contained either A or B or both. For instance, in the IBPS System a search request for information on Mark Twain must include *Clemens* as a search authority word. Thus a search for the logical sum of *Twain* and *Clemens* would retrieve both Wector, Dixon, *Sam Clemens of Hannibal* and Wiggins, Robert A., "Mark Twain and the Drama." [N.B. In actual practice such a measure may sometimes be unnecessary. Both of these items were indexed under the surname subject-heading *Twain* in MHRA.] It is also possible (and frequently will be necessary) to search for the logical sum of more than two search authority words. One case in point is that of subject-headings which cover the same or overlapping subject matter, but which differ in name. For example, the category "Seventeenth and Eighteenth Centuries" in *PMLA* is nearly the same as "1607-1800" in *American Literature*, while MHRA uses two categories ("17th-Century" and "18th-Century") to cover the period, and *DAL* and *YWES* include the overlapping categories "Puritanism" and "Colonial" respectively. A search for the logical sum of these six subject-headings would retrieve the items listed under each of them in the source bibliographies.

The fact that no vocabulary control as such is exercised over the data will make necessary a search for the *logical sum* of a number of terms whenever a single search authority word (other than a surname subject) is requested. For example, a search for items dealing with *adages* should include as search authority words at least *aphorism*, *axiom*, *byword*, *cliche*, *maxim*, *motto*, *platitude*, *proverb*, *saw*, and *saying*, as well as the various forms of each word (e.g., *aphorist*, *aphorists*, *proverbial*, etc.).[4] A search on this basis would retrieve every entry which contained any one of these search authority words in its subject-headings, title, abstract, or annotation.

The matter of vocabulary control will be taken up in the section "An IBPS Thesaurus." For the sake of simplicity in the following explanations, let us assume that all searches for words associated with any primary search authority word (as *aphorism*, *axiom*, etc., are associated with *adage*) are effected automatically.

LOGICAL PRODUCT is a function which requires the co-occurrence (or coordination) of two or more terms. That is to say, searches for any two or more search authority words co-occurring in any single bibliographic entry will be possible. For example, a search for the logical product of *Faulkner* and *Town* would retrieve, among others, the following items:

---

[4] The capability, possessed by the IBPS program EUREKA, of conducting "open-ended" searches obviates the necessity of including in the Search Authority more than one form of a given word. For instance, a search on APHORIS...would retrieve all instances of *aphorism*, *aphorisms*, *aphorist*, *aphorists*, etc.

Faulkner.[5] Harder, Kelsie B., "Proverbial Snopeslore," *TFSB*, XXIV, 89-95. [*The Hamlet* and *The Town.*][6]

Lombardo, Agostino, "*The Town* di William Faulkner," *Criterio*, II, 132-139.

Or, one might raise the number of search authority words to three and search on *Henry* (to distinguish him from the rest of his family), *James*, and *bibliography* in order to retrieve the titles of bibliographies devoted to Henry James. The result would include:

Anon., "The Soho Bibliography of Henry James," *CLQ*, Ser. iv, pp. 277-279.[7]

Edel, Leon, and Dan H. Laurence, *A Bibliography of Henry James*, London: Rupert Hart-Davis, 1957; Fair Lawn, N.J.: Essential Books.

Nowell-Smith, Simon, "Without Benefit of Bibliography: Some Notes on Henry James," *BC*, VII, 64-67.[7]

LOGICAL COMPLEMENT is an "other than" function which is of negligible importance to the IBPS retrieval program except in combination with other functions. The bibliographic answer to a request for "all entries *other than* those labelled 'Miscellaneous'" or for "all entries *other than* those on Melville" would be an enormous, meaningless welter of items. But when coupled with *logical sum* and/or *logical product*, logical complement can be an important help to the user. For example:

Sum + Complement: Items on (A) *Emerson* and/or *Thoreau*, except those which appeared in (Z) *ESQ* [(A + B) - Z]; items on (C) *Twain* and/or *Clemens*, except those on (Y) *Following the Equator* [(C + D) - Y].

---

[5] Surname subject-heading from *PMLA*.

[6] Annotation recorded from *PMLA*.

[7] These items may or may not be James Bibliographies.

Product + Complement: Items on (E) *Myth* and (F) *Symbol*, except those written by (X) Leslie Fiedler [(E · F) - X]; items on (E) *Myth* in (G) *Hemingway*, but not (W) *archetype* [(E · G) - W].

Two other combinations are, of course, possible and valuable. The most necessary will be:

Sum + Product: e.g., (*James + Howells + Twain + Clemens + Crane*) · *Realism*.

And all three functions can be combined in one operation: e.g., (*Naturalism + Realism + Slice-of-life*) · *Norris - Dreiser*.

2. An IBPS Thesaurus. In dealing with any information system over which no vocabulary control has been exercised from the outset (i.e., prior to input), two rather serious problems must be faced and solved: (a) false drops--the retrieval of irrelevant entries, and (b) lost items--the failure to retrieve relevant entries. The first of these, false drops, may not prove to be quite so serious a problem to the IBPS System as may at first appear, for unless the number and nature of the false drops deceives expectations, a staff member or the user himself can readily identify and eliminate them from any one-subject bibliography. Despite the fact that some control is, in reality, exerted over a large part of the data in the form of surname subject-headings (similar to Calvin Mooers' "Identifiers"), the present uncontrolled state of the indexing vocabulary is clearly intolerable (although the degree of intolerability cannot be predicted with much precision).

Until the IBPS System possesses a thesaurus, any user who requests a list of works on "Whitman's metrics" will certainly not receive a complete list from the file if he simply asks for a search for the logical product of *Whitman* and *metrics*. His request will not, for instance, cause the retrieval of entries indexed under the surname subject-heading *Whitman* which contain such words as *versification*, *meter*, *prosody*, etc., rather than *metrics*, in their titles or predicative matter. The same holds for nearly any coordinate search which might be proposed: the scattering of similar and identical subjects under a wide range of differing subject-headings, key words in titles and annotations, etc., makes the choice of search authority words for the retrieval of most single-subject bibliographies a very complex matter. A search for the occurrence of any search authority word other than a surname subject-heading must necessarily include, as coordinate (logical sum) search authority words, a number of synonyms and related words in order to ensure retrieval of all relevant items.

The process of listing such words for each search as it is requested would obviously be expensive and wasteful of staff time. It seems, therefore, necessary to construct, in an early stage of any continuation of this Pilot Study, a thesaurus, an associative listing of index terms, to cope with this problem.

Preliminary research shows that the most practicable means of constructing a thesaurus capable of alleviating the problem is by manually conflating selected terms from the following sources:

1. Thrall, W. F., and Addison Hibbard. *A Handbook to Literature*, revised and enlarged by C. Hugh Holman. New York: The Odyssey Press, 1960.

2. Duffy, Charles, and Henry Pettit. *A Dictionary of Literary Terms*, rev. ed. New York: Brown Book Company, 1953.

3. The Merriam-Webster *Webster's Dictionary of Synonyms*.

4. Lewis, Norman, ed. *The New Roget's Thesaurus of the English Language in Dictionary Form*. Garden City, N.Y.: Garden City Books, 1961.

5. Any good desk dictionary which makes a practice of listing and discriminating synonyms.

6. The index to Spiller, Robert E., et al. *Literary History of the U.S.* New York: The MacMillan Company, 1962, adds a few valuable entries (though their synonyms must be sought in sources 1-5).

Each of these sources has its limitations and (sometimes serious) shortcomings, but in combination they may very well provide the resources for a thesaurus which will adequately meet the needs of an information service for literary scholars. (Where these sources obviously fail, or leave detectable lacunas, the deficiencies would, of course, have to be filled by the ingenuity of the researchers compiling the thesaurus.)

Optimum exploitation of such an aid will probably be attained in one of two ways (or possibly in some combination of the two): (a) by maintaining the thesaurus on magnetic tape in the computer with an automatic lookup function in the programming which would cause the inclusion as coordinate search terms (logical sum) in the search authority of all words (synonyms, associated words, etc.) listed with the original search authority word in the thesaurus; (b) by maintaining the thesaurus in printed form so that, on receiving a search request, a staff member could choose from the thesaurus the search authority words most likely to satisfy the request.

A further point should not be ignored: although such a thesaurus will be valuable and necessary to the Integrated Bibliography for the next several years, the existence of an adequate thesaurus would ultimately make possible a machine assignment of standardized index terms (an accomplishment whose main purpose would be to reduce machine search time).

III. CONCLUSION

This Pilot Study has demonstrated, it is believed, that an Information Service for English Studies based on an Integrated Bibliography is feasible.

The incompleteness of every single bibliography in the field and the advantages to be gained--in accessibility of information, coverage, and accuracy--by integrating these bibliographies are obvious and were thus the *raison d'etre* of the Pilot Study. The fundamental problems, therefore, were not whether bibliographies could be integrated, but by what *efficient means* their integration could be attained; and whether, and how, such an Integrated Bibliography could be utilized to produce an effective Information Service.

It was assumed that the most efficient and worthwhile means of achieving an Integrated Bibliography for English Studies would be one which required an absolute minimum of human editorial activity and a consequent substitution of the computer to carry out editorial functions. It should be possible to input data to the computer in its original form, and to carry out by computer program the analysis of this data requisite to its integration and retrieval. The IBPS System was designed to meet these and other important criteria. Its implementation has been carried far enough to demonstrate that the criteria have been met. Specifically, the system:

(1) Will accept a *literatim* transcription of each source bibliography via paper tape, punched cards, or (ultimately) magnetic tape produced by optical scanning equipment, preserving all elements of typography and all predicative matter in relation to its proper Bibliogram(s).

(2) Will adapt by computer operations the unedited bibliographical source material to the process of automatic searching and automatic merging, by identifying the elements of each bibliographic entry and attaching to each element a distinctive machine-sensible tag.

(3) Provides a Master File which is amenable to searches under a wide variety of combinations of Search Authorities and Rubrics, such selective searches being necessary to the full exploitation of the inherent retrieval capability of the data.

(4) Will accept additional data of several kinds: new entries, corrections and additions to entries already lodged in the Master File, or entire new bibliographies.

(5) Is capable of producing output in elegant, as well as simple, typography, in any desired format.

Some of the far-reaching implications of this study should also be mentioned. Though the system designed during this study was created for a specific purpose, it clearly has other applications. Because the problems encountered are common to the processing of bibliographical data of many kinds, their solutions apply to the handling of bibliographical data in general. Thus the IBPS System directly relates, on the one hand, to information services in other disciplines, and on the other hand to the information-handling and dissemination problems of libraries and other repositories of data whose records are in "bibliographical" form.

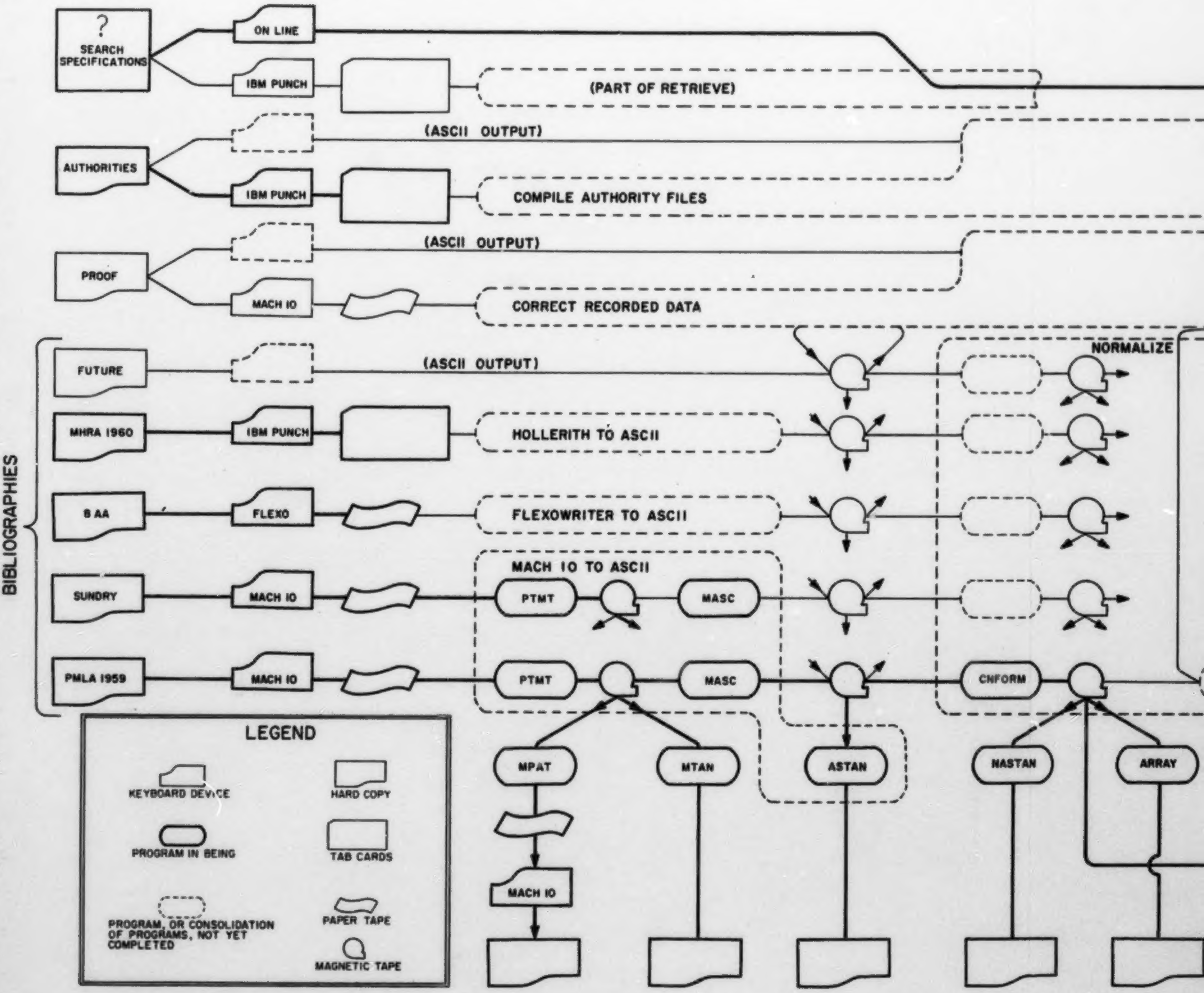
Second, the capability for sophisticated output (via photocomposition or computer-driven typesetting equipment, or suitably-equipped high-speed printer) makes possible the production of typographically attractive books. The most relevant applications of this capability are to the publication of bibliographies, although the system is adaptable to other kinds of text. For example, it would seem to be reasonable that the manuscript texts of forthcoming serial bibliographies be run through the system, thus updating the Integrated Bibliography Master File and also setting each text in publishable form.

In view of these several advantages found in the IBPS System, it is recommended that implementation of the system be completed so that (a) production-type operations can be carried through on relatively large quantities of data, to the end that (b) the retrieval capability of the system can be thoroughly tested, and that (c) cost estimates for the operation of the system as an Information Service can be obtained. When these steps are carried out, it will then be appropriate to inquire whether, when, and by whom such an Information Service for English Studies should be operated. It would seem that a cooperative enterprise involving professional organizations in the field, the institution where the system was developed, and the sponsor of the present study would be most desirable.

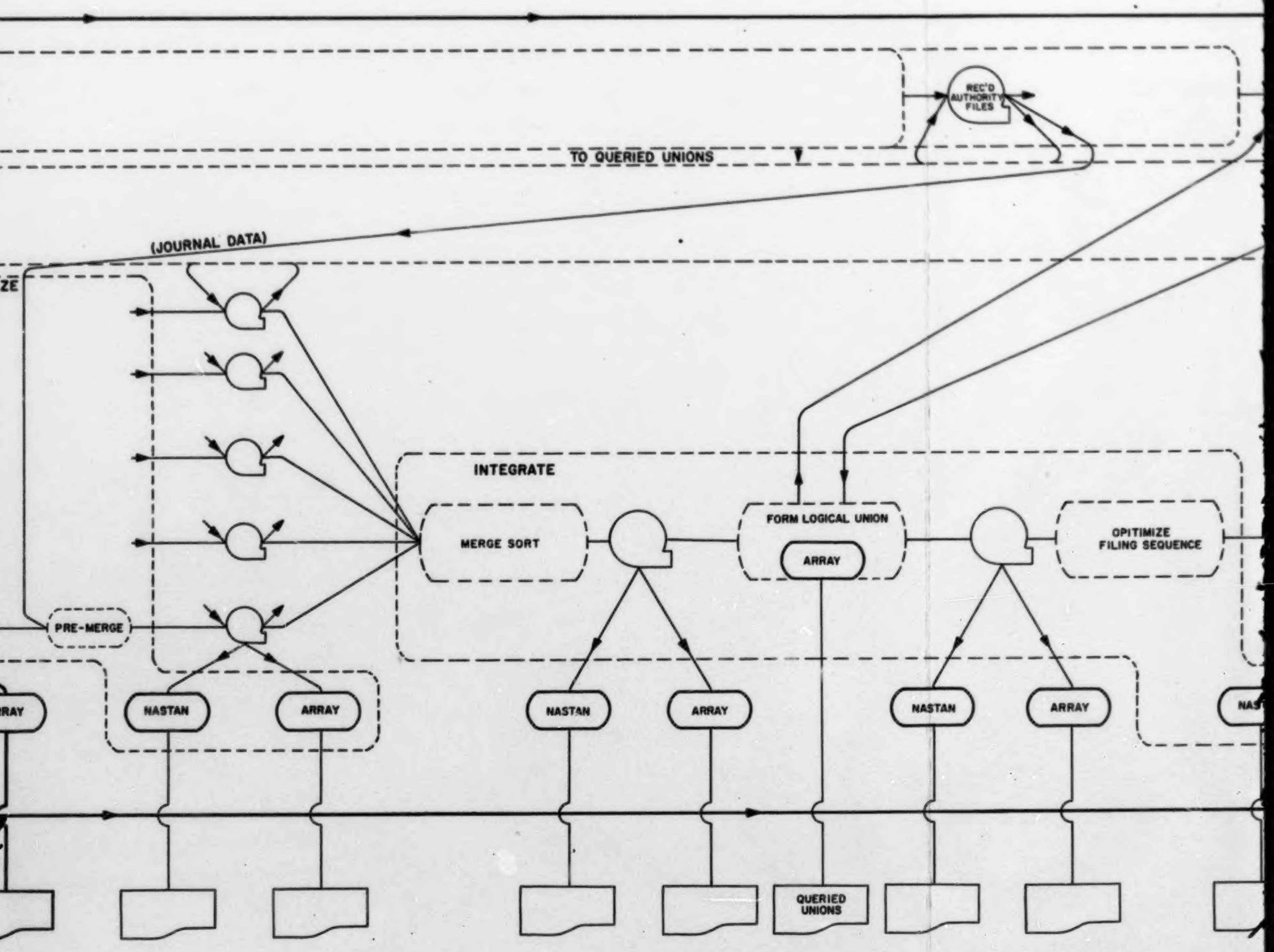
APPENDIX A

DIAGRAMS

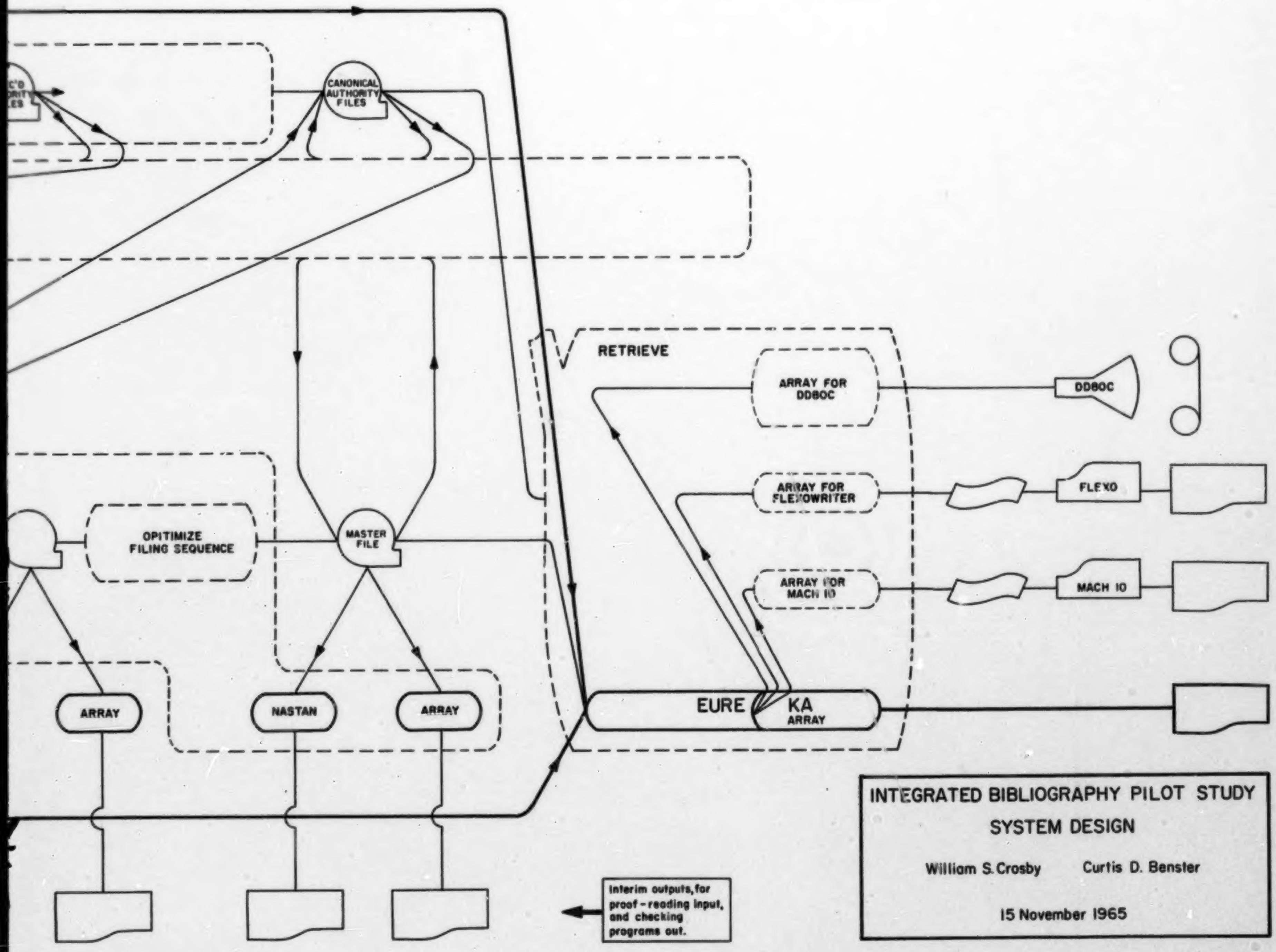
Section 1


?
SEARCH SPECIFICATIONS
ON LINE
IBM PUNCH
(PART OF RETRIEVE)
AUTHORITIES
(ASCII OUTPUT)
IBM PUNCH
COMPILE AUTHORITY FILES
PROOF
(ASCII OUTPUT)
MACH IO
CORRECT RECORDED DATA
BIBLIOGRAPHIES
FUTURE
(ASCII OUTPUT)
NORMALIZE
MHRA 1960
IBM PUNCH
HOLLERITH TO ASCII
BAA
FLEXO
FLEXOWRITER TO ASCII
SUNDRY
MACH IO
MACH IO TO ASCII
PTMT
MASC
PMLA 1959
MACH IO
PTMT
MASC
CNFORM
MPAT
MTAN
ASTAN
NASTAN
ARRAY
MACH IO
LEGEND
KEYBOARD DEVICE
HARD COPY
PROGRAM IN BEING
TAB CARDS
PROGRAM, OR CONSOLIDATION OF PROGRAMS, NOT YET COMPLETED
PAPER TAPE
MAGNETIC TAPE

Section 2


REC'D AUTHORITY FILES
TO QUERIED UNIONS
(JOURNAL DATA)
ZE
INTEGRATE
MERGE SORT
FORM LOGICAL UNION
ARRAY
OPITIMIZE FILING SEQUENCE
PRE-MERGE
RAY
NASTAN
ARRAY
NASTAN
ARRAY
NASTAN
ARRAY
NAS
QUERIED UNIONS

Sec. 3


C'D
ORITY
LES
CANONICAL
AUTHORITY
FILES
RETRIEVE
ARRAY FOR
DDBOC
DDBOC
ARRAY FOR
FLEXOWRITER
FLEXO
ARRAY FOR
MACH 10
MACH 10
OPITIMIZE
FILING SEQUENCE
MASTER
FILE
ARRAY
NASTAN
ARRAY
EURE
KA
ARRAY
Interim outputs, for proof-reading input, and checking programs out.

INTEGRATED BIBLIOGRAPHY PILOT STUDY
SYSTEM DESIGN

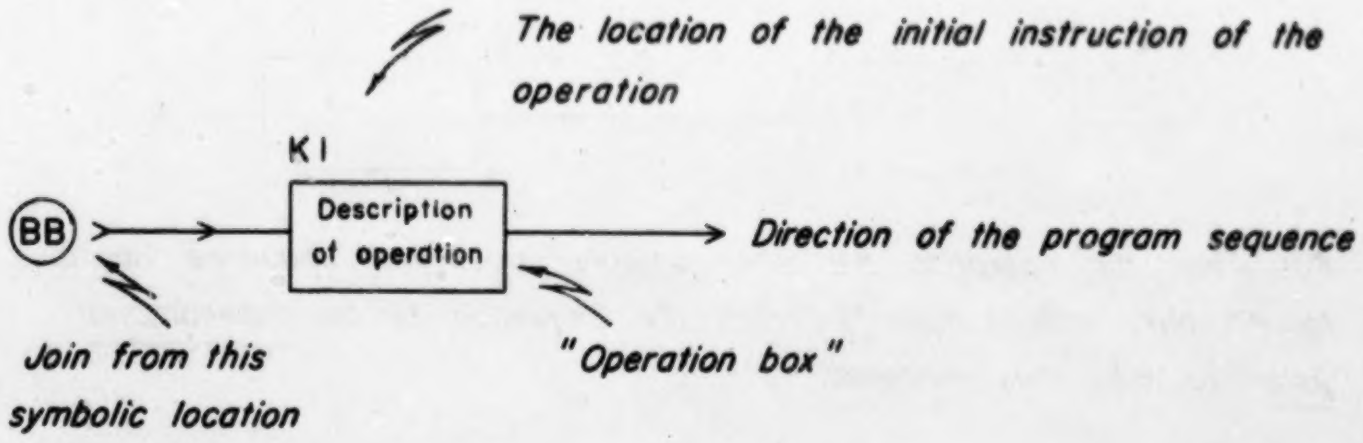
William S. Crosby　Curtis D. Benster

15 November 1965

INTEGRATED BIBLIOGRAPHY PILOT STUDY

LEGEND TO PROGRAMS

In the logical diagram ("flow diagram"):

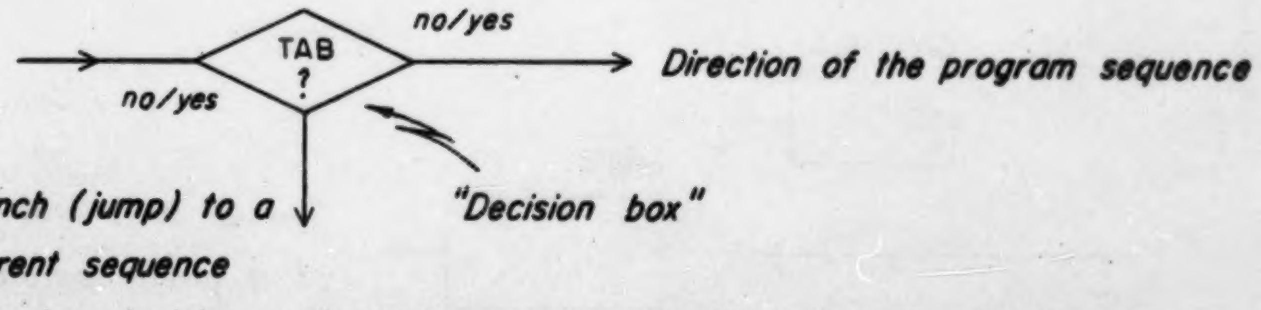


The location of the initial instruction of the operation

Direction of the program sequence

Join from this symbolic location

"Operation box"



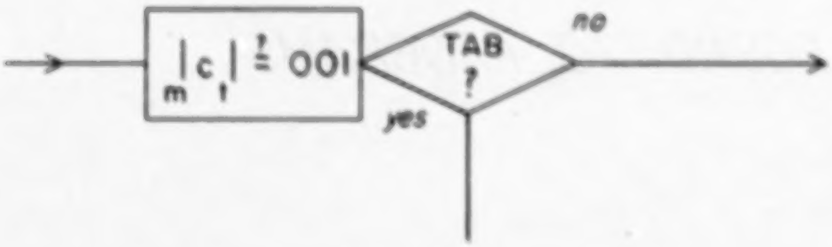
Direction of the program sequence

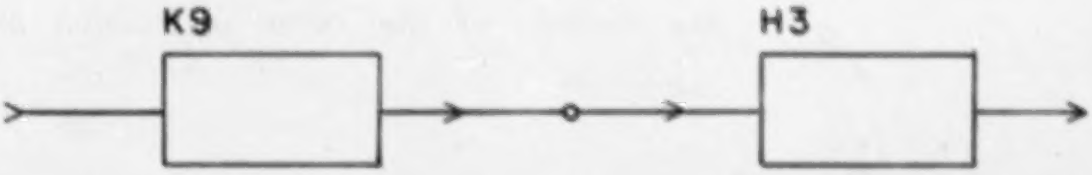
Branch (jump) to a different sequence

"Decision box"

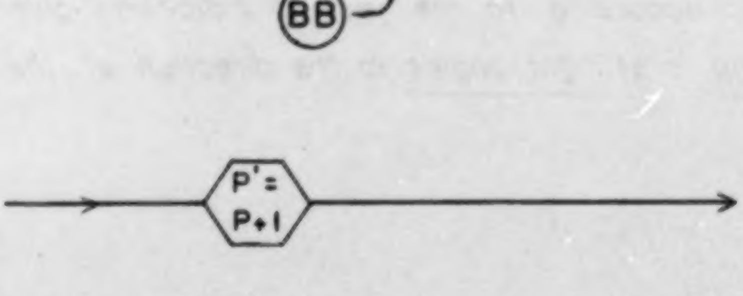
Following decision, the program pursues either its normal sequence, or jumps to a different one, according to the yes/no indication given on the respective line. (The jump is at right angles to the direction of the program sequence.)

The specific numerical (or logical) test that must be applied, and the specific data to which it is applied, are often given previous to the "decision box" in an attached "operation box" (see also below under "notation"):

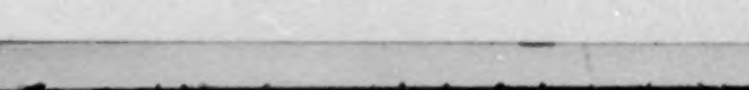




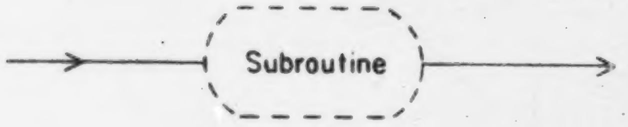
Following the operation K9 the computer does not sequence normally to H3 but, instead departs from the sequence by an unconditional jump to H3, thus indicated.



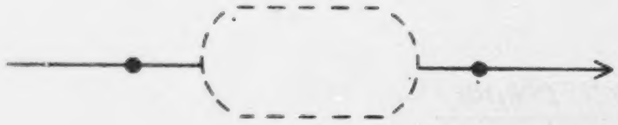
Several lines may join at a given operator.



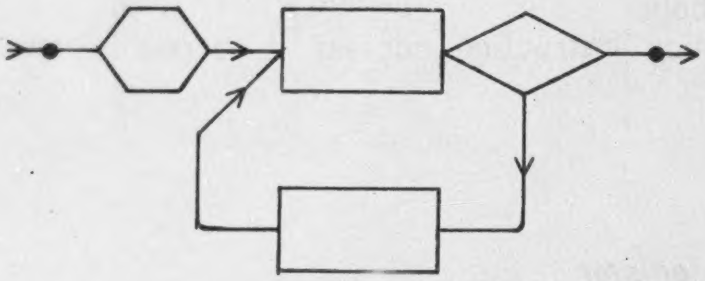
For operations upon an index the hex is reserved.

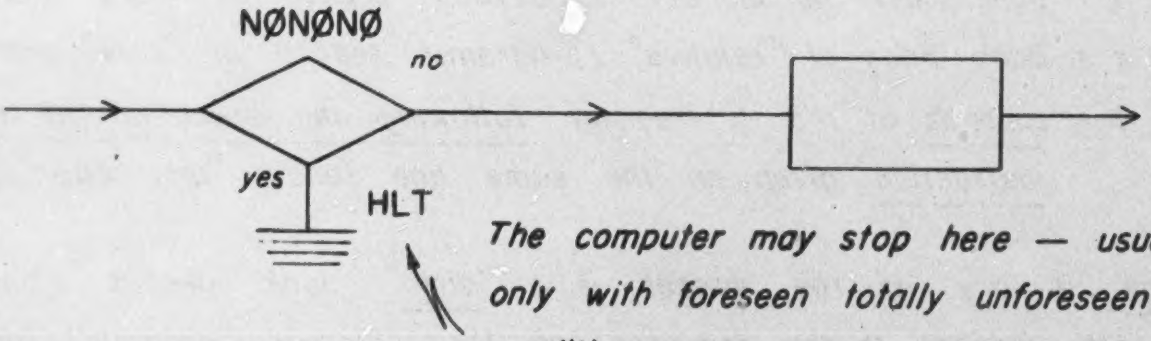


A subroutine, common to many lines of the logic, may be referred to.

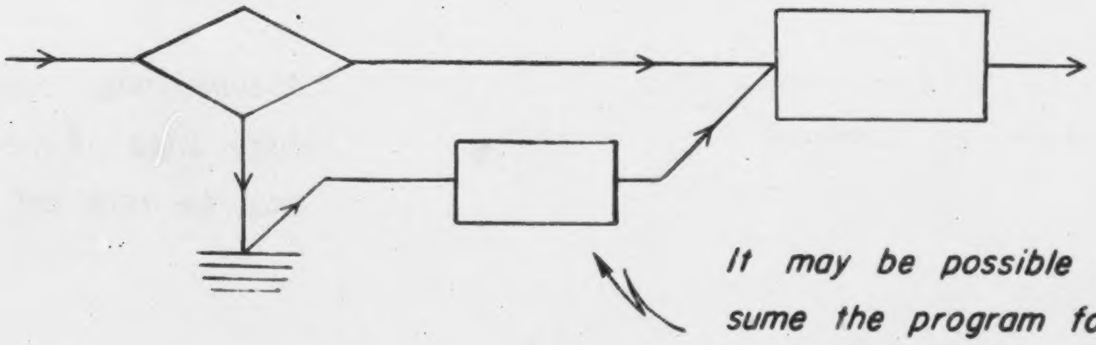


When such a subroutine has been reached by a "jump and record" order, the black dots (jumps with record) indicate this; they are useful also to indicate the bounds of a section of main program utilized thus:





The computer may stop here — usually only with foreseen totally unforeseen conditions.



It may be possible to resume the program following such a stop.

In the coding sheet (computer program proper):

The resulting contents of all relevant registers are cited alongside every computer instruction. (N.B.) So is each change in storage. E.g., in the CDC-160A coding sheets, columns exist, to the left of the coded program proper, giving:

b d i r	A	Storage		Symbolic location	Instruction	Operand address	Δ address	Comments

where

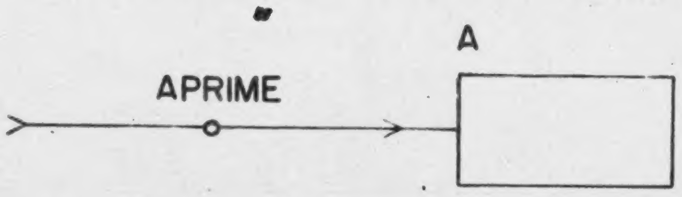
b = bank index of buffer register
d = bank index of "direct" section of "core" storage
i = bank index of "indirect [indexed] section of "core" storage
r = bank index of "relative" [program] section of "core" storage
A = contents of the A register following the execution of the instruction given on the same line (under "instruction")

A blank in any of the spaces is a "ditto": look upwards (backwards with respect to the sequence of the computer program) for the required information.

Following a break in the sequencing of the program register, i.e., after a (factually) unconditional jump instruction, a double line bars the reader from continuing in sequence:

Symbolic location	Instruction	Operand address		
APRIME	JFI	I		
		A		

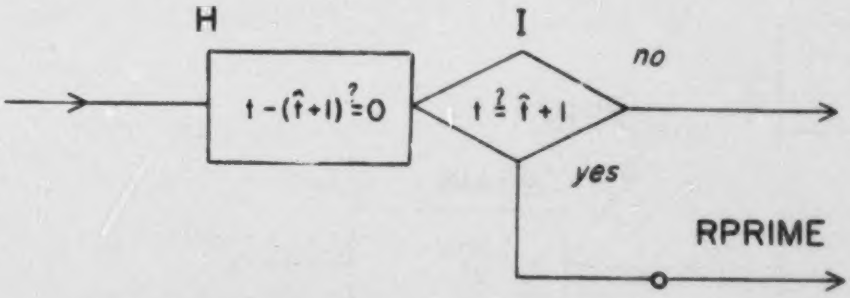
Thus this will in general appear on the flow chart as:



Following a (factually) conditional jump ("branch") order, on the other hand, is a single line:

			Symbolic location	Instruction	Operand address		
A t $t-(\hat{t}+1)$			H	SBD	TMAX+1		
			I	ƵJF	RPRIME		
$_m\|c_t\|$				LDI	T		

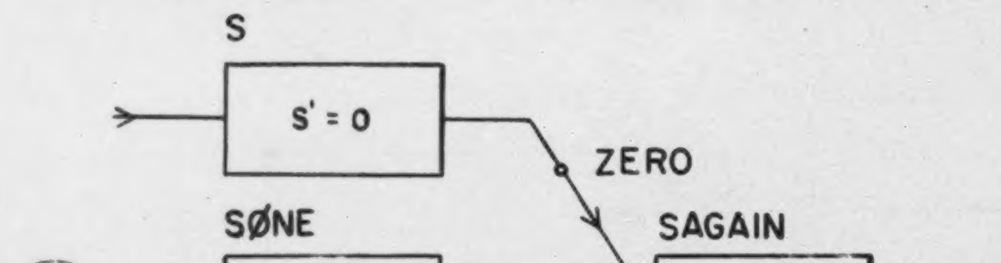
and this conforms with the branch off the main line of the program (i.e., out of the proper sequence) at a "decision box:



Nota bene: the *line at right angles* to the preceding line of the flow diagram coincides with the *branch* of the instruction. If the decision box is not a single conditional jump order but the consequence of a resolution by subroutine, however, this proviso is dropped. Owing to the presence of double lines, the "ditto" (blank) in the register accounting must be terminated — and afterwards resumed by a re-statement of the register contents — at the break in the program sequence thus indicated. The additional space between successively listed — but not successively sequenced — instructions provides for this "carry forward" in the accounting:

bdir	A	Storage	Joins from	Symbolic location	Instruction	Operand address
010						
	S'=0			S	LDN	00
				ƵERØ	ƵJF	SAGAIN
010			AGØ			
	S'=1			SØNE	LDN	Q1
010	S'		ZERØ			
		S'		SAGAIN	STI	[S]

A special column is reserved for "joins from". Like the carry-forward of the data in the register accounting (which often must be given in half-height letters, as here, to make the continuity legible), the source of the join is cited in the space above the joined instruction. Bracketed data are subject to modification in the course of the program. The coincidence of coded program and flow diagram is thus complete; for the above coded program we have:



= = = = =	*Computer halt; sequencing resumable.*
▬ ▬ ▬ ▬ ▬	*Transfer to a subroutine via a record order.*

These few very simple conventions, virtually self-explanatory, are observed throughout the Integrated Bibliography programming by the author (C.D.B.); they afford the strongest guarantee against the programming's being bound by any particular computer.

Notation:

The notation has been developed with a view to making clear, not snowing!

c	*a character:* h, \$, c, CR, TAB, ☐, *etc.*
$\lvert c\rvert$	*the numerical value of that character (some combination of 7 bits, say).*
$_m\lvert c\rvert$	*the numerical value of that character in Mach code.*
$_{an}\lvert c\rvert$	*the numerical value of that character in Anelex code.*
$\underline{\overline{062}}$	*the character whose numerical value is 062.*
$\lvert c_t\rvert$	*the character (appearing as its numerical value), drawn on the index* t.

For example, if the storage location of our character is 0000 + 7 = 0007, then with the index t *set to 0007, this character can be properly located and, say, brought into the* A *register for subsequent operation:*

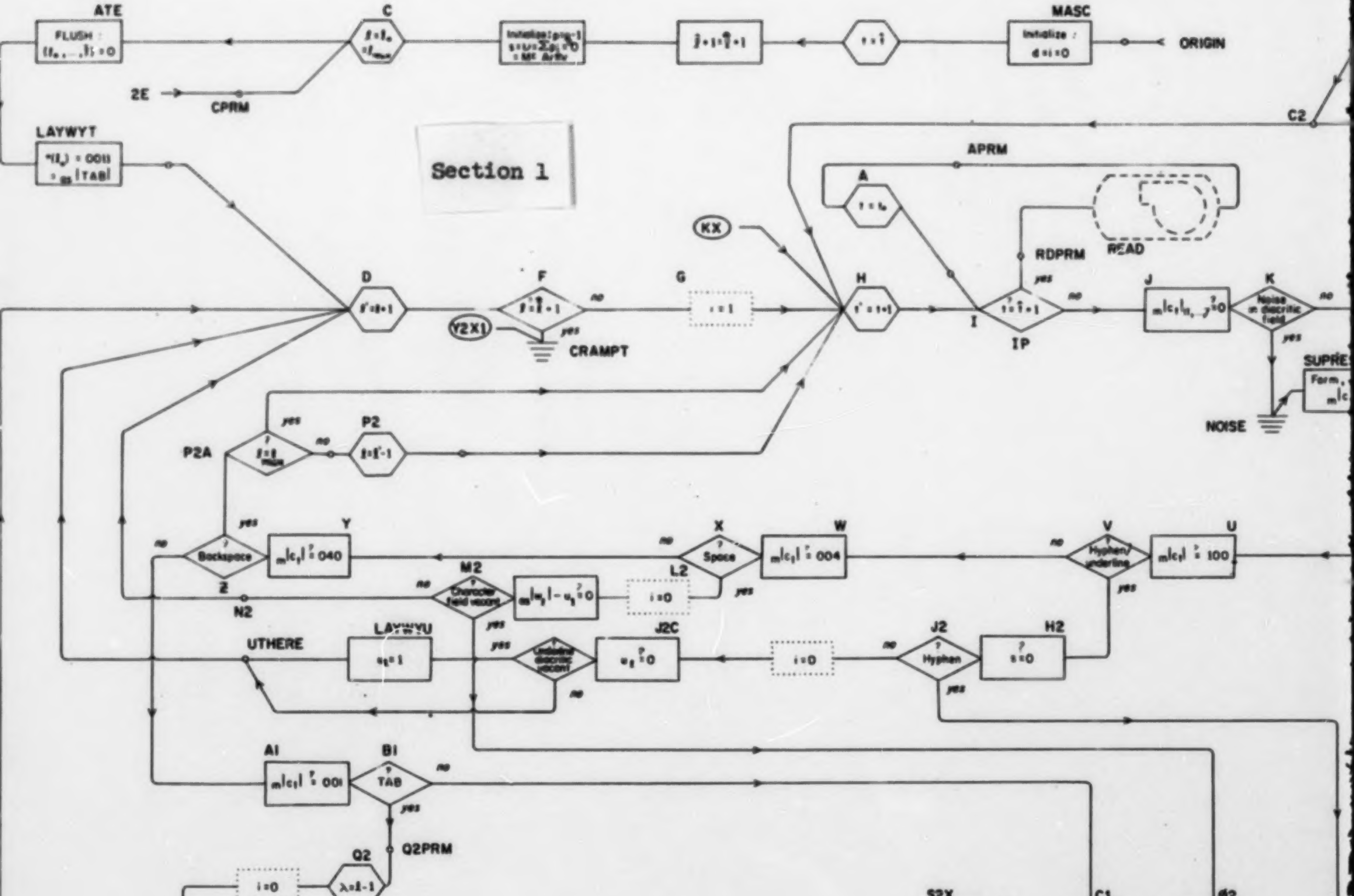
b d i r	A	Storage	Joins from	Symbolic location	Instruction	Operand address
I						
	$_m\lvert c_t\rvert$				LDI	T

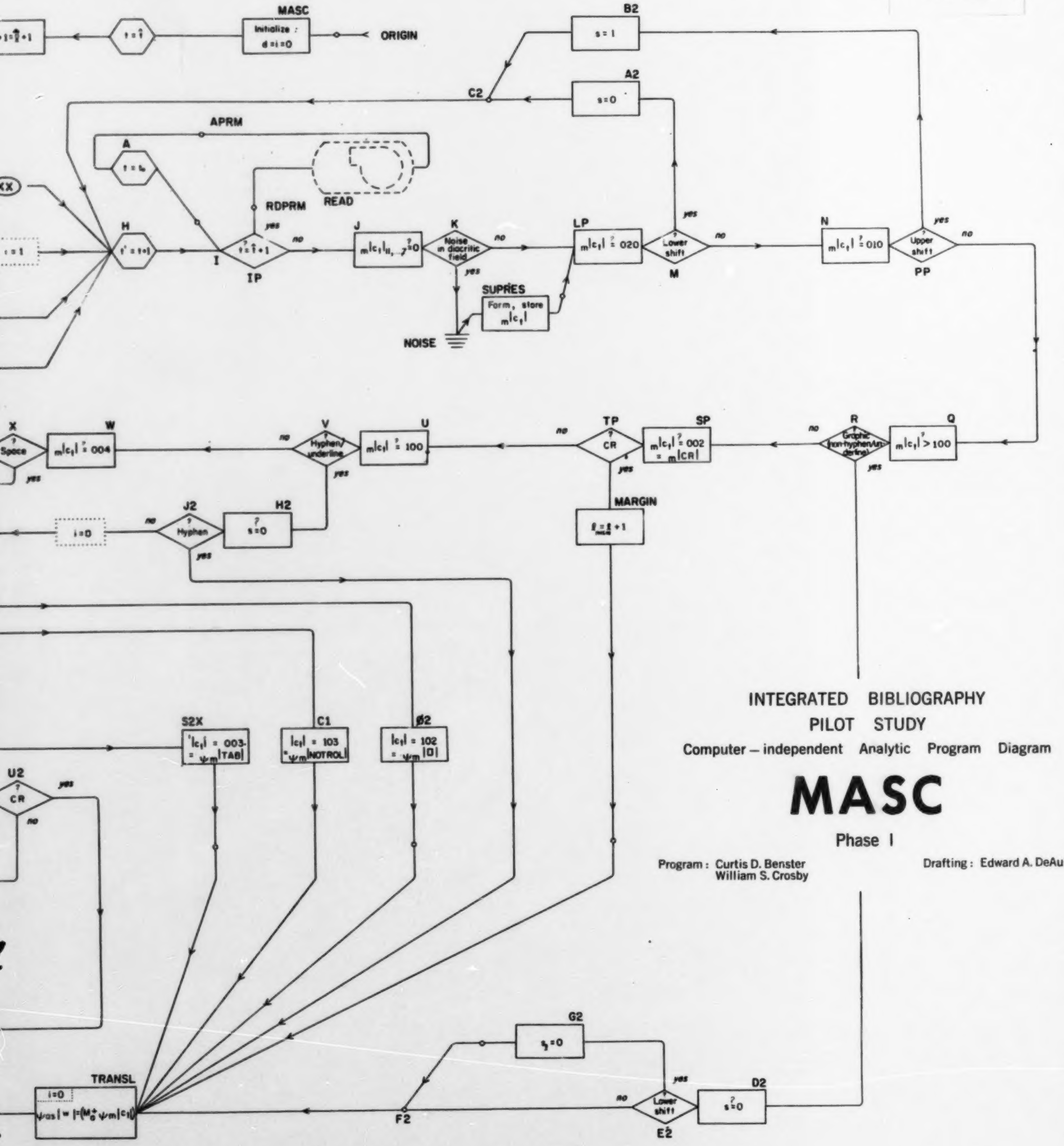
${}_m\|c_t\|_6$	*The bit in position 6 [counting as the computer does, starting with the bit in position 0 at the right] of the numerical value of the character specified.*
$t, p, \ell,$	*Indexes. In the operation of the CDC-160A computer they include their base, say*
t_0	*the initial value of the index. t_0 may be 0; but p_0 may be 1200, say, as in one program.*
$\hat{t}$	*the maximum value attained by the index t*
$\hat{t}+1$	*the unattained upper bound of the index t, one more than the attained maximum.*
t'	*the new value of the variable t; ordinarily $t' = t + 1$, that is, the new value equals one more than the old value. (after this incrementing, the variable is again called indiscriminately t)*
(...)	*the contents of ...*
)...(	*the location of ...*
λ	*this is a copy (an "image") of the index ℓ; such an image may be modified for searches, etc., without affecting the proper value of ℓ — upon which the stability of the indexing depends.*
τ, π	*similarly the image of t and p respectively.*

The supreme facility of modern mathematics — where it is facile indeed and not self-obstructed — owes an incalculable debt to supreme felicities of notation, often the choice of centuries of use. The above notation, applicable to the characters of written natural language, is offered only as the fruit of a winnowing already not inconsiderable. Like the preceding conventions, these are calculated to be both readily manipulable and readily intelligible.

Curtis D. Benster Boulder, Colorado 19 November 1964

Section 1


ATE
FLUSH :
C
MASC
Initialize :
d=i=0
ORIGIN
2E
CPRM
LAYWYT
C2
APRM
A
READ
RDPRM
KX
D
F
G
i=1
H
I
IP
J
K
Noise in diacritic field
no
yes
Y2X1
CRAMPT
SUPRES
NOISE
P2A
P2
Y
X
Space
W
$_m|c_1| \overset{?}{=} 004$
V
Hyphen/underline
U
$_m|c_1| \overset{?}{=} 100$
Backspace
$_m|c_1| \overset{?}{=} 040$
Z
N2
M2
Character field vacant
L2
i=0
UTHERE
LAYWYU
J2C
J2
Hyphen
H2
AI
$_m|c_1| \overset{?}{=} 001$
BI
TAB
Q2PRM
Q2
i=0
S2X
C1
Ø2
$|c_1| = 103$
$= \psi_m|NOTROL|$
$|c_1| = 102$
$= \psi_m|D|$
R2
$_{as}|w_\lambda| \overset{?}{=} 0011$
S2
T2
$_{as}|w_\lambda| \overset{?}{=} 0015$
U2
CR
LAYWYR
Y2
Y2X
Y2X1
CRAMPT
V2
X2
MASC II
LAYWY+1
(ℓ)^5777
LAYWY
Preserve $\psi_{as}|w|$
TRANSL
i=0
$\psi_{as}|w| = (M_0^+ \psi_m|c_1|)$
JUMP
F2


MASC
Initialize : d = i = 0
ORIGIN
t = t̂
t̂ = t̂ + 1
B2
s = 1
A2
s = 0
C2
APRM
A
t = t₀
READ
RDPRM
yes
H
t' = t + 1
I
t ≟ t̂ + 1
IP
no
J
m|c₁|₁₁,...₇ ≟ 0
K
Noise in diacritic field
no
yes
SUPRES
Form, store m|c₁|
NOISE
LP
m|c₁| ≟ 020
Lower shift
M
yes
no
N
m|c₁| ≟ 010
Upper shift
PP
yes
no
Q
m|c₁| ≟ > 100
R
Graphic (non-hyphen/underline)
no
yes
SP
m|c₁| ≟ 002 = m|CR|
TP
CR
no
yes
MARGIN
ℓ = ℓ + 1
U
m|c₁| ≟ 100
V
Hyphen/underline
no
yes
W
m|c₁| ≟ 004
X
Space
yes
H2
s ≟ 0
J2
Hyphen
no
yes
i = 0
S2X
|c₁| = 003 = ψm|TAB|
C1
|c₁| = 103 = ψm|NOTROL|
Ø2
|c₁| = 102 = ψm|0|
U2
CR
yes
no
TRANSL
i = 0
ψas| = |=(M₀⁺ ψm|c₁|)
F2
G2
s₂ = 0
Lower shift
E2
yes
no
D2
s ≟ 0

INTEGRATED BIBLIOGRAPHY
PILOT STUDY
Computer – independent Analytic Program Diagram

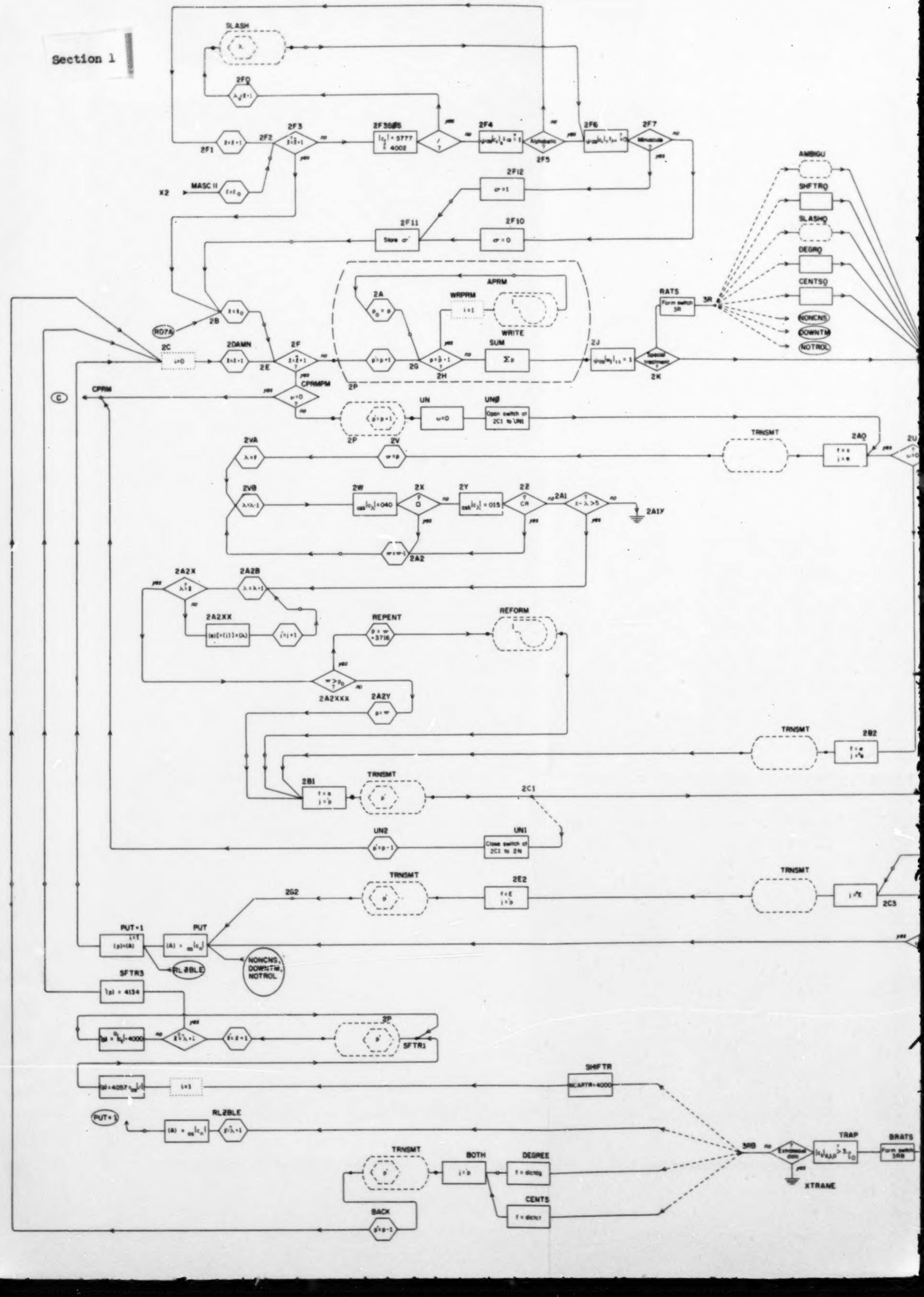
MASC

Phase I

Program : Curtis D. Benster
William S. Crosby

Drafting : Edward A. DeAuila

Section 1



SLASH
2FO
2F1
2F2
2F3
2F3&5
4002
2F4
2F5
Alphabetic
2F6
2F7
Minuscule
yes
no
X2
MASC II
2F12
σ = 1
2F11
Store σ'
2F10
σ = 0
AMBIGU
SHFTRQ
SLASHQ
DEGRQ
CENTSQ
NONCNS
DOWNTM
NOTROL
RATS
Form switch 3R
3R
APRM
2A
WRPRM
WRITE
SUM
Σp
2J
Special treatment
2K
RO7A
2B
2C
i = 0
2DAMN
2E
2F
2G
2H
2P
CPRMPM
CPRM
C
UN
ω = 0
UNØ
Open switch at 2C1 to UN1
TRNSMT
2AO
2U
2VA
2V
2VB
2W
2X
2Y
2Z
CR
2A1
2A1Y
2A2
2A2X
2A2B
2A2XX
REPENT
REFORM
2A2XXX
2A2Y
2B2
2B1
2C1
UN2
UN1
Close switch at 2C1 to 2N
2G2
2E2
2C3
PUT+1
PUT
RLZBLE
NONCNS, DOWNTM, NOTROL
SFTR3
(p) = 4134
SFTR1
SHIFTR
PUT+1
3RB
Extended data
TRAP
BRATS
Form switch 3RB
XTRANE
BOTH
DEGREE
CENTS
BACK

INTEGRATED BIBLIOGRAPHY
PILOT STUDY
Computer – independent Analytic Program Diagram

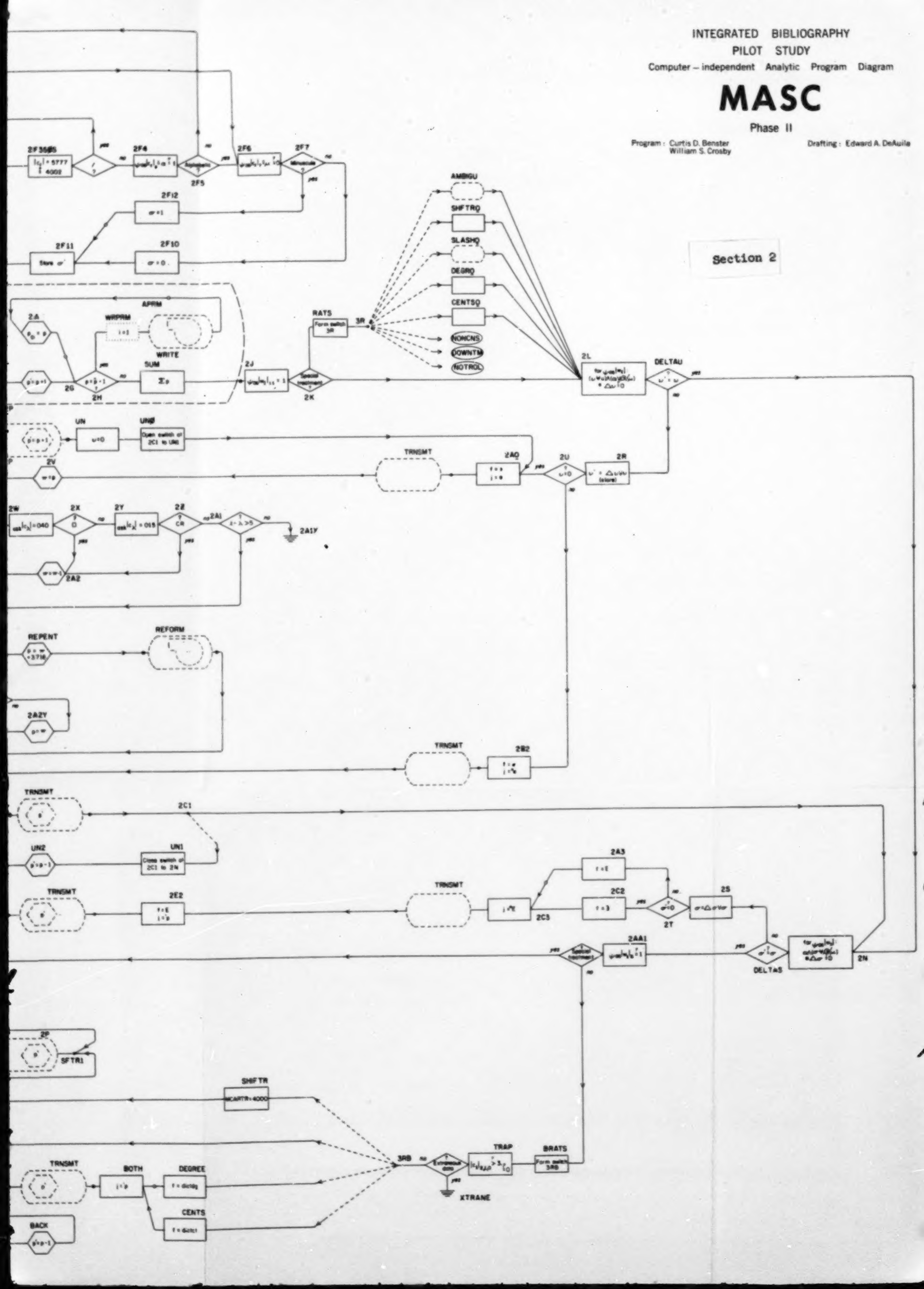
MASC

Phase II

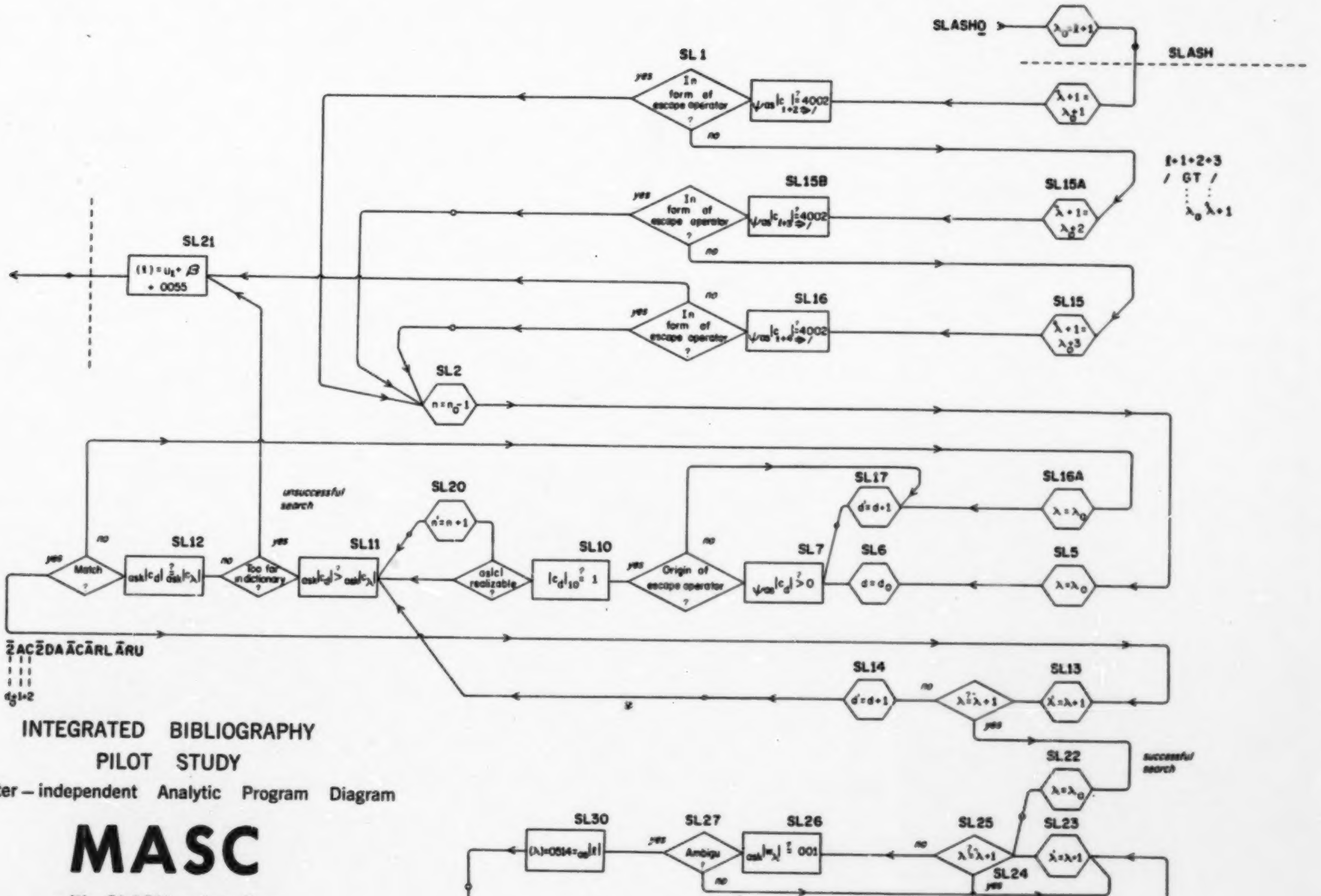
Program: Curtis D. Benster
William S. Crosby

Drafting: Edward A. DeAuila

Section 2



2F3585
2F4
2F5
2F6
2F7
2F12
2F11
Store σ'
2F10
$\sigma = 1$
$\sigma = 0$
AMBIGU
SHFTRQ
SLASHQ
DEGRQ
CENTSQ
NONCNS
DOWNTM
NOTROL
RATS
Form switch 3R
3R
APRM
WRPRM
WRITE
SUM
Σp
2A
2G
2H
2J
2K
Special treatment
2L
DELTAU
yes
no
UN
UNØ
Open switch at 2C1 to UN
TRNSMT
2AQ
2U
2R
(store)
2V
2W
2X
2Y
2Z
2A1
2A1Y
2A2
CR
REPENT
REFORM
2A2Y
2B2
2C1
UN2
UN1
Close switch at 2C1 to 2N
2E2
2A3
2C2
2C3
2T
2S
2AA1
DELTAS
2N
2P
SFTR1
SHIFTR
TRAP
BRATS
Form switch 3RB
3RB
Extraneous data
XTRANE
BOTH
DEGREE
CENTS
BACK


SLASHO
SLASH
SL 1
In form of escape operator ?
yes
no
SL15B
SL15A
SL21
SL16
SL15
SL2
n = n0 - 1
f+1+2+3
/ GT /
λ0 λ+1
SL20
n' = n + 1
unsuccessful search
SL12
SL11
SL10
SL17
d' = d + 1
SL16A
λ = λ0
Match ?
Too for in dictionary ?
ascii realizable
Origin of escape operator ?
SL7
SL6
d = d0
SL5
ZAC ZDA ACARL ARU
d0 1+2
SL14
SL13
successful search
SL22
SL30
SL27
Ambigu ?
SL26
SL25
SL23
SL24
SL22A
SL22AA
SL23A
Realizable inserter ?
SL24A
SL25A
Modifier ?
SL26A
SL27A
α-inserter ?
SL30A
SL31
i.e., shifter
SL42
SL43
Ascii control ?
SL44
SL55
B = 0
SL46
(λ)' = B + 4002
SL52
λ + 2
SL52A
SL47
(λ)' = B + 0134
SL50
SL53
(λ)' = B + 0057
SL35
SL36
Terminal slash ?
SL37
SL40
SL51
SL33
B = 0200
SL34
(λ)' = B + 4001

INTEGRATED BIBLIOGRAPHY
PILOT STUDY

Computer – independent Analytic Program Diagram

MASC

with SLASH subroutine

gram : Curtis D. Benster
William S. Crosby

Drafting : Edward A. DeAuila

APPENDIX B

THE INTEGRATED BIBLIOGRAPHY PILOT STUDY
--TECHNICAL RATIONALE

Curtis D. Benster
Technical Director

To merit its name "Pilot Study," this project had to incorporate, in a working system, a very wide range of technical innovations.

These innovations are calculated to solve, in a durable way, the problems of bibliographic merging and retrieval, by electronic devices, in conditions of electronic technology so rapidly progressing as to threaten even well conceived systems with sudden obsoleteness.

The *viability* of the IB system, then, has been of prime concern: each step in the fulfillment of current responsibilities has been so taken as to guarantee the possibility of long-range future development as well.

Other innovations are designed to escape the circumscriptions of prevailing practice in the handling of library and, indeed, literary data in natural language. Thus we have considered it imperative, first, to dispense with the well established internal computer codes called binary-coded decimal, together with all similar codes, which vary from one to the other in accordance with the vagaries of the manufacturer of the respective piece of equipment. This is veritably to cast off an incubus. But it entailed dispensing with the software developed over the last twelve years or so, the effects of which self-denial were of course felt in the early stages of our work. To adequately represent just the punctuation and like signs of the printed data of our sources required such a code as the proposed American Standard Code for Information Interchange (ASCII), then forthcoming, and at the outset the decision was made to pioneer a use of that code. Very many other desiderata were attained by use of it, not least the prospect of alphabetizing on the basis of its rational ordering of the English alphabetics and the numerals.

Not less imperative was the transcending of the limitations of this new code forthwith, for it yielded but a longer tether, however refreshing the sense of freedom conferred by first use of it. No compromise was to be made in the representation of the diacritics of the Western European languages. These are customarily omitted in computer treatment; but this is illiteracy, offensive to the educated user, and vitiating search and retrieval. An entirely new system of *escape operations and symbols* was accordingly developed (William S. Crosby); while remaining consistent with ASCII and even with its international counterpart from the International Standards Organization, this very greatly generalizes typographical shifting of font, making it possible for us to maintain exact records from a range of fonts practically infinite, and to print fittingly from these records when the appropriate printing equipment evolves in the future. Yet input can remain a function of the present cramped keyboards, for the escape symbols are very simply mnemonic, and indeed lend themselves to accurate insertion by unskilled keyboard operators, our experience shows.

All details of typography are, then, preserved, to a nicety. But this now makes possible the most telling technical revolution of all: the processing of the matter of a published bibliography *without any pre-editing whatsoever*: it is left to our computer program to determine, from criteria of typography and format--very minute criteria but, taken together, sufficient--just what is *author*, what is *title*, what is *"predicative"* (as we call it, i.e., classificatory), what is *collation*, etc., in the bibliographic record. The "information separators" provided by ASCII are inserted by computer program: thus we have turned our backs not only upon pre-editing, but upon the time-honored fixed fields of IBM card lore; the record may be a continuum of whatsoever length, and thus the transition from bibliography to abstract and, indeed, text, is inherently guaranteed by the IB system.

The assignment of these information separators by computer program solely is an accomplished fact: working on our verbatim transcript of a bibliography of very great complexity (*Publications of the Modern Language Association*), we have succeeded in making this assignment accurately, later exploiting it in searches over certain rubrics only--title of book but not title of article, for instance, or surname subject but not author.

This identification by computer program of the proper recorded information, accompanied by the assignment of the appropriate information separators, is called by us a "normalizing" process: by it a record from any bibliography can be reduced to a uniform tractable form. The adaptability of the normalizing process to bibliographies of differing format and typographical characteristics is guaranteed by elaborating it in smaller subroutines which differ from bibliography to bibliography only by sequence and parametric content.

An implication of this achievement is that our system is technically equal to the verbatim input of an *optical reader* whensoever the latter is equal to the multilingual text of the Modern Language Association. We are in a position to turn our backs again upon immured practice, upon the tedious and institutionalized re-typing or re-keying of an existing competent record, for the IB system need depend only upon the verbatim record of optical reader or, as well, compositor's original.

Certain standards have been held to throughout the work which further guarantee the capacity of the present work to serve as a solid foundation for further development in many directions.

All basic analyses have been embodied in analytical program diagrams, so formulated as to be independent of the specifics of any computer. But there has been no recourse to "common" or "intermediate" "languages" for the programming (e.g., Fortran); our programming has been in virtual machine language, subject to transformation by elementary assembly only. It is correspondingly compact and forthright. It will be possible with ease to interpret the perspicuous analytical program diagrams when making future modifications of the programs to accommodate new data; it will be equally easy to re-program in the language of a new computer when one becomes available or when it is desired to utilize the different computers of other institutions for the execution of the Integrated Bibliography system.

To further guarantee the intelligibility of the system, the documentation includes a program scroll for each program, strictly conformable with the analytical program diagram, on which the detail of each operation is given algebraically, together with a running record of register contents and changes of storage. The notation is covered by a *Legend* and will occasion no difficulty to a newcomer.

In the manipulation of the data, character-by-character operation has been followed scrupulously, making possible the iterative treatment of very irregular mass data. String operations are the culmination of such a processing, not its basis, and thus a great simplification in the overall treatment has been achieved.

A small (desk-size) computer (the excellent CDC-160A) has been used, by choice, for the initial execution of the system. Owing to its own high speed and logical power, and especially to the power of its associated equipment (magnetic tape transports, high-speed printer, paper-tape reader and high-speed punch, high-speed card reader)--and of course the essential tractability of the character-by-character processing--processing speeds, including overall search speeds, have been attained that are on a par with those yielded by much larger computers. The practicability of the system for libraries, traditionally impecunious, is thus further guaranteed.

* * *

A means of evaluating the potential of the Integrated Bibliography system has now put in its appearance at the NCAR computing facility many months before it could have been hoped for by an unassisted library project--an ultra-high-speed printer, the DD80, utilizing a cathode-ray tube and a high-speed camera.

It will be possible presently to utilize this output device to print in some typographical detail the result of successful searches of the Integrated Bibliography--or for that matter, the entire Bibliography. Majuscule and minuscule letters and italics--the latter necessary to distinguish title within title, for example--as well as bold-face fonts, can be realized without appreciable further programming; an output in arbitrary array, say one resembling a Library of Congress Card, will then be available at extraordinary speeds, on the order of a thousand lines a second.

The deeper accuracy and power of the Integrated Bibliography record is, however, in its diacritics and shifts to foreign fonts--Cyrillic, or Arabic, or for that matter Devanagari. Such scripts the DD80 will trace by vector, which requires both the design of type fonts and the appropriate DD80 programming.

And it is first of all in this direction that the potential of the IB system ought now to be explored.

APPENDIX C

THE 1964 IBPS CONFERENCE

The Conference on the Integrated Bibliography Pilot Study was held at the University of Colorado, in Boulder, on July 29-31, 1964. The objective of the Conference was to obtain suggestions and criticism on the work done and projected in the Pilot Study, and to seek guidance on certain elements of the later phases--notably retrieval. In order to achieve this objective, it was necessary

1. to present a description of the Pilot Study--its aims, scope, and progress;
2. to outline some possibilities for the application of modern information-system techniques to bibliography in English Studies;
3. to provide a forum for comment by Conference participants.

These needs structured the Conference program, in which roughly the first half was devoted to presentations on points (1) and (2), the last half to discussion sessions. The Proceedings of the Conference may be found in *The IBPS Progress Report*, pp. 1-85.

The Conference, which brought together English scholars, librarians, information retrieval and computer specialists, and representatives of government and industry, was, it is felt, highly successful. While fewer suggestions than anticipated were received on the specifics of retrieval, there was a correspondingly greater amount of criticism on the other phases of the Study--criticism which has been carefully considered and evaluated.

Perhaps the most interesting discovery made in the Conference was that the IBPS system for encoding data in machine-readable form, which requires very little human editorial effort and a correspondingly large employment of the computer in "editorial" functions, is both relatively new, and controversial. For an experimental study such as the IBPS, this is as it should be. There is, after all, no doubt that bibliographical entries can be made machine-searchable, or that several bibliographies can be conflated by human effort. What is important is to find out (1) to what extent human editorial functions can be replaced, and (2) what kinds of information can be derived from an integrated bibliographical record.

CONFERENCE ON THE INTEGRATED BIBLIOGRAPHY PILOT STUDY

July 29-31, 1964
University of Colorado
Boulder, Colorado

ROSTER OF PARTICIPANTS

Jeremiah M. Allen
English in Engineering
Univ. of Colorado
Guggenheim 106
Boulder, Colorado

H. F. Arader
IBM
Scientific & Technical
Information, Corporate
Armonk, New York 10504

George Arms
Univ. of New Mexico
607 Vassar Dr. NE
Albuquerque, New Mexico 87106

Lt. Col. C. L. Balch
Department of English
USAF Academy, Colorado

Eric H. Boehm
American Bibliographical Center
800 E. Micheltorena Street
Santa Barbara, California

John Broderick
Library of Congress
4718 Albemarle St., N.W.
Washington, D.C. 20016

Mrs. Lillian V. Cottony
National Bureau of Standards
Boulder Laboratories Library
Boulder, Colorado 80301

Don S. Culbertson
Colo. State Univ. Library
Fort Collins, Colorado 80521

R. P. Davis
Air Force Accounting &
Finance Center
3800 York Street
Denver, Colorado 80205

Vinton A. Dearing
Univ. of Calif. at Los Angeles
Los Angeles, Calif. 90024

Seymour L. Flaxman
State Univ. of New York at
Stony Brook
Stony Brook, New York 11790

Steve Furth
Data Processing Division
IBM Corporation
112 East Post Road
White Plains, New York 10601

Donna Gerstenberger
Department of English
University of Washington
Seattle, Washington 98105

Eugene Graziano
Lockheed Corporation

Murry Grigsby
Jefferson County Public Schools
1580 Yarrow Street
Lakewood, Colorado 80215

Harold A. Haswell
U.S. Office of Education
400 Maryland Ave. S.W.
Washington, D.C. 20202

Phoebe F. Hayes
Bibliographical Center for
Research
Rocky Mountain Region, Inc.
c/o Denver Public Library
1375 Broadway
Denver, Colorado 80203

Muriel Hoppes
Law Library, Univ. of Colo.
Fleming Law Bldg.
Boulder, Colorado

Gerald B. Kauvar
1029 S. 8th Avenue
Denver, Colorado 80218

Lewis Leary
Dept. of English & Comparative
Literature
Columbia University
Philosophy Hall
New York, New York 10027

J. Robert Lebsack
National Bureau of Standards
Boulder Laboratories
Boulder, Colorado 80301

Lawrence E. Leonard
National Bureau of Standards
Boulder Laboratories Library
Boulder, Colorado 80301

Jack M. McCormick
National Center for
Atmospheric Research
Boulder, Colorado

Vincent J. McGee
Air Force Accounting &
Finance Center
3800 York Street
Denver, Colorado 80205

Austin J. McLean
Humanities Librarian
Colorado State Univ. Library
Fort Collins, Colorado 80521

Roy M. Mersky
Univ. of Colo. Law Library
Fleming Law Bldg.
Boulder, Colorado

Lt. Col. L. M. Meier
Air Force Accounting &
Finance Center
3800 York Street
Denver, Colorado 80205

Col. P. R. Moody
Department of English
USAF Academy, Colorado

Eleanor Nicholes
Harvard College Library
22 Chauncy Street
Cambridge, 38, Massachusetts

Charlotte W. Ratcliffe
National Bureau of Standards
Boulder Laboratories Library
Boulder, Colorado 80301

Robert F. Richards
Univ. of Denver
Denver, Colorado

F. B. Rogers
Univ. of Colo. Med. Center
1155 Ash Street, Apt. 201
Denver, Colorado 80220

Edward Rose
Dept. of Sociology
Univ. of Colorado
Boulder, Colorado

I. Rotkin
Harry Diamond Laboratories
Washington, D.C. 20438

Gertrude E. Schalow
Bureau of Reclamation, Denver
2584 S. Gaylord
Denver, Colorado 80210

Donald A. Sears, Exec. Sec.
The College English Assoc.
2401 H Street, NW, Apt. 505
Washington, D.C. 20037

Marjorie B. Stephenson
Hd. of Reference, Ariz. State Univ.
1701 Cutler Dr., #C
Tempe, Arizona

Laurence M. Thorsheim
Air Force Accounting &
Finance Center
3800 York Street
Denver, Colorado 80205

Arlin Turner
Duke University
1115 Woodburn Road
Durham, North Carolina

Donald R. Tuttle
U.S. Office of Education, DHEW
4th & Maryland
Washington, D.C. 20202

James Woodress
San Fernando Valley State College
Northridge, California

T. W. Ziehe
The RAND Corp.
1700 Main Street
Santa Monica, California

CONFERENCE ON THE
INTEGRATED BIBLIOGRAPHY PILOT STUDY
July 29-31, 1964
University of Colorado
Boulder, Colorado

PROGRAM

July 29

8:30 - 9:00 a.m.	University Memorial Center, Rm. 336	Registration
9:00 - 10:15 a.m.	University Memorial Center, Rm. 336	"The Integrated Bibliography and the Pilot Study: A Review of Assumptions and Objectives." (Charles Nilon and Lewis Sawin)
10:30 - 12:00 n.	University Memorial Center, Rm. 336	"The Pilot Study: A Progress Report." (Mr. Sawin, William Crosby, Gene Cunnar, Phil Loney, Allen R. Penner)
12:00 - 1:00 p.m.	Faculty Club 972 Broadway	Luncheon. ($1.25. Please indicate at registration if you will attend.)
1:30 - 2:00 p.m.	NCAR and Computing Center	Tour of University Computing Center and NCAR facilities.
2:00 - 4:30 p.m.	NCAR and Computing Center	Demonstration of the bibliography. (Mr. Crosby) Question Period.
8:00 - 10:00 p.m.	University Memorial Center, Rm. 336	"Application of Indexing Techniques to an Integrated Bibliography of English Studies." (Jessica Melton, Western Reserve University Center for Documentation and Communication Research; Roger Clark, IBPS Staff.)

July 30

9:00 - 10:15 a.m.	University Memorial Center, Rm. 336	Discussion session: "Requirements of an Information System for English Studies, I."
10:15 - 11:00 a.m.	Norlin Library Rm. 401	Coffee for participants (courtesy Norlin Library Staff Assn.)
11:00 - 12:15 p.m.	University Memorial Center, Rm. 336	Discussion session: "Applications of the Integrated Bibliography."
1:15 - 2:45 p.m.	University Memorial Center, Rm. 336	Discussion session: "Requirements of an Information System for English Studies, II."
3:00 - 4:30 p.m.	University Memorial Center, Rm. 336	Discussion session: "Requirements of an Information System for English Studies, III."
5:30 p.m.	Home of Mr. & Mrs. Charles Nilon, 2005 Spruce Street	Reception for participants.

July 31

9:30 a.m.	University Memorial Center, Rm. 336	Discussion session: "Phase II of the Pilot Study--which Approaches?"

APPENDIX D

SPECIAL STUDIES

QUESTIONNAIRE

Prior to the Conference, the following questionnaire was distributed to approximately 100 scholars and teachers in American literature. Its purpose was to discover what type of bibliographical classification system would be preferred by actual users of bibliographies. A tabulation of the results of this questionnaire may be found in *The IBPS Progress Report*, pp. 79-81.

We are making a pilot study in mechanized bibliographical retrieval in American literature and are concerned about an adequate classification and indexing system for American literature bibliography. Perhaps such a system already exists. Perhaps the modification of some existing system, which has in it certain flaws, might produce an adequate system. Perhaps a new and adequate system can be devised. Your answers to the questions below may help us in thinking clearly and purposefully about the problems in this area. If you can, and care to, please answer the question(s) below that you feel you can answer. Thank you.

1. What is your special field of interest?
2. What bibliographical sources do you consult most often when you work in your field?
3. Under what subject headings do you usually look for material when you consult your usual sources?

4. Would you be happier if the subject headings in the tools you use were different?

   What or how would you like them to be?

5. Indicate if you can which of the tools that you use often has the best subject headings?

6. Point out any tools that use subject headings which you consider inadequate.

   What suggestions can you give to make these tools adequate?

7. Are you especially disturbed when different indexers place scholarly article under different headings?

8. What kinds of questions should a scholar ask about classification and indexing systems?

Please add to this list of questions any other relevant questions of the same kind whose answers you feel we should be concerned with.

INDEXING STUDIES

It was generally felt that, prior to any commitment to an indexing scheme for the IBPS, a detailed study of modern indexing methods should be undertaken. The first tangible fruits of this study were three papers delivered at the 1964 IBPS Conference, on Coordinate Indexing, KWIC Indexing, and Citation Indexing, and their relation to the IBPS (see *IBPS Progress Report*, pp. 46-72). While such special kinds as the KWIC and citation indexes proved tempting, it was readily concluded that they lay too far outside the immediate aims of the IBPS to be considered for implementation during the Pilot Study phase of the Integrated Bibliography. Likewise, special kinds of coordinate indexes (e.g., TABLEDEX) are a temptation because of their neatness, but a retrospective bibliography of the size contemplated by the IBPS simply cannot be accommodated by them without the expenditure of vast sums of money and great human effort. Even a Uniterm index, it was decided, is not readily adapted to use on such a data base.

The value of the study of modern indexing, then, was not that it led to the adoption of an existing indexing technique (nor was it much expected that it would), but that, by making known what had been and was being done, the study greatly influenced and directed the conception of indexing and retrieval in the IBPS system.

"A SELECTED BIBLIOGRAPHY ON INFORMATION RETRIEVAL"

As an aid to the directors and research personnel of the Integrated Bibliography Pilot Study, a bibliography on information retrieval was compiled which has proved most helpful to them, though it does not at all attempt to be exhaustive.[1] In order to help the participants of the 1964 IBPS Conference orient themselves toward information retrieval and its problems by directing their attention to further readings in the field, the IBPS working bibliography was reduced to "A Selected Bibliography on Information Retrieval" and distributed to all conferees. It contains 340 items in these categories:

General
Bibliography
Reference Works
Information Theory
Indexing and Abstracting: General
Coordinate Indexing
Citation Indexing
Computers and Linguistic Problems
Selective Dissemination of Information
Systems Testing and Evaluation
Hardware
Automated Information Techniques

Though the reception of the bibliography was good--even, in some instances, enthusiastic--a retrospective look indicates that, though selective, the bibliography was perhaps not sufficiently selective for the purposes of the Conference. Many conferees, it developed, had extensive knowledge of the literature in the field, while others had no acquaintance with the subject at all. Since the Conference was conducted primarily for the latter, it appears that a greater need existed for a much more selective bibliography which omitted many of the technical reports and items on such subjects as "Hardware," and which was, perhaps, annotated, than for a bibliography which attempted to serve, at once, two so divergent groups.

---

[1] For an exhaustive listing of books and articles in the field, see Charles F. Balz and Richard H. Stanwood, *Literature on Information Retrieval and Machine Translation* (Owego, N.Y.: IBM Corporation), November, 1962.

In this reduced, dittoed form, however, the bibliography has continued to be a more useful tool for the IBPS staff than was the full bibliography of some 700 items which existed only on 3x5 cards.

Dittoed copies of the bibliography are available from the directors.

"A BIBLIOGRAPHY OF BIBLIOGRAPHIES OF AMERICAN LITERATURE"

"A Bibliography of Bibliographies of American Literature" is an outgrowth of long range planning for the Integrated Bibliography Pilot Study. In the sense that the study is a pilot or feasibility study, it is essential to have not only a system for bibliographical retrieval, but also to know to some extent the quantity and character of the bibliography to be retrieved. Because the trial bore is in books and articles on American literature, it was deemed proper to study first the character and quantity of bibliography in American literature. To discover the extent of American bibliography a survey was made of selected standard bibliographical tools of the sort contained in Mr. Gohdes's guide to the study of American literature. In the initial phase, entries were taken from:

Gerstenberger and Hendrick, *The American Novel 1789-1959.*

Leary, *Articles in American Literature.*

Modern Humanities Research Association--Annual Bibliography, 1920-1961.

Publications of the Modern Language Association, "Annual Bibliography," 1922-1964.

Spiller, et al., *Literary History of the United States.*

Stovall, *Eight American Authors.*

The result was a reasonable concept of scope and, as well, the core of a working bibliography of bibliographies in American literature. With this data in hand, a second group of sources was examined:

Blanck, *Bibliography of American Literature.*

Woodress, *Dissertations in American Literature.*

Gohdes, *Bibliographical Guide to the Study of the Literature of the United States.*

Quinn, *The Literature of the American People.*

Thurston, et al., *Short Fiction Criticism.*

Trent, et al., *Cambridge History of American Literature.*

Trubner, *Guide to American Literature.*

A very substantial basic bibliography of bibliographies of American literature--approximately 2,500 entries--resulted. As the corpus developed, it became increasingly evident that what had been begun as an attempt to determine scope was evolving into a scholarly tool in its own right. Since it is believed that there is not now in existence a bibliography of bibliographies in American literature (though such a tool was at one time proposed), the decision was made to expand.

Expansion has taken three forms: increasing the body of entries, standardizing their form, and arranging them in a meaningful and useful order.

(1) If a bibliography of this sort is to be a useful tool to scholars, it must be as comprehensive as possible. Accordingly, efforts were made to tap all possible sources of bibliography. Libraries and presses specializing in this kind of output were contacted for their lists. Sample volumes from Sabin, *A Dictionary of Books relating to America* (29 vols.), Evans, *American Bibliography* (14 vols.), and *The Library of Congress Catalogue* (310 vols.) were surveyed to determine the length of time involved and the approximate yield in entries which could be expected from each. Literary histories, sale and exhibition catalogues, and various periodicals have been examined. The result is a body of approximately 4,000 entries, with a potential for as many as 6,000.

(2) Since entries were taken from a variety of sources, they vary greatly in form, accuracy, and completeness. When the interest was only scope, such variety presented no problems; however, when the aim became the provision of a scholarly tool, standardization of form became essential. A form was agreed upon, and the process of authentication and formalization of entries is near completion.

(3) Considerable thought has been given to a method of classification and cross-indexing which will make the bibliography most readily useful to scholars. Sample classifications have been constructed with this aim in mind. From this work and planning a truly serviceable classification system will evolve.

The University of Colorado Press has expressed a strong interest in publishing the completed bibliography of bibliographies in American literature, but a final decision has not yet been made. Should this arrangement fail to materialize, a publisher will be sought elsewhere.

COST ESTIMATES

The IBPS has not yet reached the stage at which cost estimates for an Integrated Bibliography may be computed.

Input. The figure now available for input costs is $8.10 per 100 entries, but that figure is suspect for a number of reasons: (a) it is an average between card punch and Dura Mach 10 punching costs; (b) it does not reflect the fact that, though two Dura machines were often in simultaneous operation, one was being supplied *gratis* by Dura Business Machines of Colorado; (c) much of the material now on paper tape has not received its final proof-reading, though these figures treat it as ready to be transferred to magnetic tape.

Output. Cost statements for output do not, of course, exist, as the IBPS has not yet achieved the practical capability of production-type output.

APPENDIX E

GLOSSARY

AUTHOR. The creator of a book, article, etc., listed in a source bibliography.

BIBLIOGRAM. A bibliographic item (*q.v.*).

BIBLIOGRAPHIC ITEM. The "title page" information given for each book, article, dissertation, etc., listed in a source bibliography, generally consisting of author, title, and publication facts. A bibliographical item *does not* include predicative matter (*q.v.*).

ENGLISH STUDIES. Studies of the literature of Great Britain and the United States, and of the English and American languages.

ENTRY. In a bibliographical source, the whole of the printed information concerning a single book, article, etc. A *Merged Entry* has, in addition to the Bibliogram, the sum of the predicative matter associated with that Bibliogram in all the source bibliographies.

ESCAPE SYMBOL. A device employed to "escape" the font limitations of the input device and to record a requisition for any character from a virtually unlimited font.

THE INTEGRATED BIBLIOGRAPHY FOR ENGLISH STUDIES. "A single bibliographic compilation containing every item which has ever been listed in a major bibliography whose subject contents are related to the discipline of English Studies, with provision for continuous addition of new items, the whole indexed and stored in such a way that retrieval of many discrete segments of the total record can be accomplished quickly and accurately on demand."

IBPS. Integrated Bibliography Pilot Study (Office of Education Cooperative Research Grant #2189).

INTEGRATION. The process of forming the qualified logical union of entries containing duplicate (though not necessarily *identical*) Bibliograms. The final product of this operation is, for each case of duplication, a *Merged Entry* containing a single Bibliogram plus the sum of the predicative matter associated with it in all the source bibliographies in which it is found.

MERGE. The integration of the entries of a source bibliography with those contained in the Master File.

NORMALIZATION. The process, accomplished by the program CNFORM, of identifying the elements of each Bibliogram so that they may be arranged in a standard order to facilitate their comparison with Bibliograms in the Master File.

PREDICATIVE MATTER. All information in an entry (or merged entry) which is not a part of the Bibliogram. Though the predicative matter varies in its composition from item to item and from source bibliography to source bibliography, it will always contain one or more of the following elements: accession (or entry) number, annotation, abstract, and classification structure (subject-headings).

SUBJECT INDEXING. A mode of indexing in which an indexer determines which words are to be used as index terms, regardless of whether they appear in the source.

SURNAME SUBJECT-HEADING. A subject-heading (given in a source bibliography) which is a "Surname Subject" (e.g., the name of a person, writer or otherwise, who is the subject of a book, article, etc.).

WORD INDEXING. Taking words directly from the source, without any change, for use as indexing terms.

The typescript of this final report was prepared by Barbara Allen on the Dura Mach 10 paper tape-punching typewriter, using an IBM Prestige Elite 72 type sphere for text and an IBM Light Italic type sphere for *italics*. The final version was typed automatically on the Dura Mach 10 by means of a punched paper tape.